Improving Single Cell ‘Omics Methods for Investigating Microbial Dark Matter

Zur Erlangung des akademischen Grades einer
DOKTORIN DER NATURWISSENSCHAFTEN

(Dr. rer. nat.)

von der KIT-Fakultät für Chemie und Biowissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
DISSERTATION
von

Morgan S. Sobol, M.Sc.

aus den Vereinigten Staaten von Amerika

1. Referentin: Prof. Dr. Anne-Kristin Kaster
2. Referent: Prof. Dr. Thomas Schwartz

Tag der mündlichen Prüfung: 15.12.2022

This document is licensed under a Creative Commons
Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0):
https://creativecommons.org/licenses/by-sa/4.0/deed.en

Declaration

This work was carried out in the working group of Prof. Dr. Anne-Kristin Kaster at Karlsruhe Institute of Technology (KIT), Institute of Biological Interfaces - 5 in the period from February 2019 to December 2022.

I hereby truthfully affirm that I have written this thesis independently, I have indicated all the aids used completely and accurately, I have marked everything that has been taken unchanged or with modifications from the work of others, and I have observed the KIT Rules for Safeguarding Good Scientific Practice as amended from time to time. Furthermore, I assure that the electronic version of this work corresponds to the written version and that the delivery and archiving of the primary data is secured at the Institute according to paragraph A (6) of the KIT Rules for Assuring Good Scientific Practice. This thesis has not been submitted in the same or similar form to any other examination authority.

Karlsruhe, December 16th 2022

Morgan S. Sobol, M.Sc.

“*We make a living by what we get, but we*
make a life by what we give”

~ Winston Churchill

Acknowledgements

First, I would like to thank my advisor Prof. Dr. Anne-Kristin Kaster for all her encouragement and guidance during my PhD. In particular, I am grateful for the numerous opportunities you have provided me with and the constant support for my success now and in the future. I look forward to our future collaborations.

Many many thanks to all of the members from IBG-5 as well as IBG-1 for assisting me during my PhD, but most importantly for your friendship. I will cherish all of the fun times we had.

Thank you as well to the BioInterfaces International Graduate School (BIF-IGS) and to the Karlsruhe House of Young Scientists (JHYS) for funding my travel to various workshops and networking opportunities.

I would like to give special thanks to my family and friends. Thank you, mom and dad for originally inspiring my curious nature of the world and for supporting my passions. And thank you to my sisters, grandparents, as well as my friends both here in Germany and abroad for your unconditional love and support.

Finally, I want to thank endlessly my husband Taylor for joining me in Germany to support my career. I will be forever grateful for your understanding and moral support during this challenging journey. This work is dedicated to you.

Zusammenfassung

Die überwiegende Mehrheit des mikrobiellen Lebens ist unentdeckt und wenig erforscht, da sie bisher noch nicht erfolgreich kultiviert werden konnten. Wir bezeichnen sie daher als mikrobielle dunkle Materie (*microbial dark matter*, MDM). MDM hat hohes biotechnologisches Potential, z.B. für die Nutzung nachhaltiger Energiequellen, zur biologischen Sanierung kontaminierter Böden, oder für medizinische Anwendungen. Der Einsatz kulturunabhängiger Methoden zur Untersuchung von Mikroorganismen in der Natur, die Metagenomik und Metatranskriptomik, hat unser Verständnis von MDM erheblich verbessert. Allerdings ist es mit diesen Methoden immer noch schwierig, einzelne Spezies bioinformatisch zu analysieren, insbesondere von Organismen mit geringer Häufigkeit in komplexen Habitaten. Stammvariationen, die falsche Zuordnung von Sequenzen, insbesondere mobiler genetische Elemente sowie sich stark wiederholende Sequenzregionen sind nur einige der Probleme, mit denen z.B. die Metagenomik konfrontiert ist. Auch bei der Metatranskriptomik führen die phänotypische Heterogenität der Zellen und die Diversität der mikrobiellen Gemeinschaften zu komplexen Transkriptionsprofilen, die nicht vollständig zugeordnet werden können. Daher wurden die Einzelzellgenomik (*single cell genomics*, SCG) und die Einzelzelltranskriptomik (*single cell transcriptomics*, SCT), die zusammen als Einzelzell-'omics (SC 'omics) bezeichnet werden, entwickelt, um die Nachteile der Metagenomik und Metatranskriptomik zu überwinden.

Die Anwendung von SCG hat sich zu einem wichtigen Instrument für die Erweiterung unseres Wissens über MDM entwickelt, beispielsweise durch die jüngste Entdeckung mehrerer neuer Phyla, von denen es derzeit nur sogenannte *single amplified genomes* (SAGs) gibt. Vollständige SAGs von vielen Mikroorganismen, insbesondere von solchen mit geringer Abundanz, sind jedoch aufgrund der vielen technischen Herausforderungen und der hohen Kosten selten. Auch die SCT ist mit den vielen Herausforderungen der Arbeit mit RNA konfrontiert, wie z. B. der kurzen Halbwertszeit von mRNA und geringen Genexpression, weshalb sie in der Mikrobiologie noch nicht häufig angewendet wird. Daher haben sich die hohen Erwartungen an mikrobielle SC 'omics noch nicht vollständig erfüllen können.

In einem typischen SCG-Arbeitsablauf können die Zellen nach der Probenentnahme vor der Einzelzellisolierung mit Fluoreszenzfarbstoffen markiert werden. Nach der Isolierung werden

die Zellen lysiert und das Genom anschließend amplifiziert, gefolgt von der Sequenzierung und bioinformatischen Datenanalyse. In dieser Arbeit wurden die Schritte der Zellmarkierung, Isolierung, Lyse und Ganzgenom-Amplifikation (*whole genome amplifikation*, WGA) verbessert, um die Methodik zu verbessern. Zunächst wurde ein Ansatz zur gezielten Zellmarkierung entwickelt, der die Anreicherung von Mikroorganismen mit geringer Häufigkeit aus Umweltproben ermöglichte. Dieser Ansatz half bei der Entdeckung neuer Phylogenien und Stoffwechseln von Mikroorganismen die in geringer Abundanz vorkommen und die andernfalls durch konventionelle Metagenomik übersehen worden wären. Darüber hinaus trägt dieser Ansatz dazu bei, die Kosten für SCG zu senken, da nun nicht mehr Zehntausende von Einzelzellen sequenziert werden müssen, um seltene Mikroorganismen zu analysieren. Als nächstes wurden die Schritte der Zellisolierung und Zelllyse verbessert, um sowohl physische Zellschäden als auch den DNA-Abbau zu minimieren, was den Erfolg des nachgeschalteten Genom-Amplifikationsschritts erhöht. Für den WGA-Schritt wurde ein Ansatz zur Volumenreduzierung systematisch getestet und etabliert, um die Homogenität und Vollständigkeit der Genomabdeckung deutlich zu verbessern. Dies Ergebnisse der Versuche zeigen, dass eine weitere Volumenreduzierung in den nL oder pL Bereich nicht erforderlich. Die Kosten der WGA konnten um 97,5 % gesenkt werden konnten, was den Durchsatz von SCG erhöhen und die Verwendung dieses Ansatzes in weiteren Forschungsgruppen positiv beeinflussen dürfte.

Da SCG allein nur Informationen über die Phylogenie, genetische Struktur und das Stoffwechselpotenzial, nicht aber über die tatsächliche Aktivität einer Zelle liefert, wurde in dieser Arbeit eine mikrobielle SCT-Pipeline entwickelt, um die individuellen Funktionen der Zelle in einer Gemeinschaft besser zu verstehen. Derzeit gibt es nur sehr wenige Methoden für mikrobielle SCT, und die, die es gibt, bleiben aufgrund ihrer schwierigen Anwendung und geringen Zugänglichkeit außerhalb ihrer jeweiligen Arbeitsgruppen weitgehend ungenutzt. Daher wurden in dieser Studie Änderungen und Verbesserungen an einer eukaryotischen Einzelzell-RNA-Sequenzierungsmethode (RNA-seq) vorgenommen, um ihre Anwendung bei Prokaryoten zu ermöglichen. Es wurde festgestellt, dass der Zusatz von Dithiothreitol (DTT) im Lysepuffer wahrscheinlich die DNase I hemmt, was zu einer DNA Kontamination führt. Die hier vorgestellten Einzelzell-RNA-seq-Ergebnisse zeigten zuverlässige Transkriptionsprofile im Vergleich zu RNA-

seq-Ergebnissen aus der gesamten Probe. Dies wurde auch durch ein *Proof-of-Principle*-Experiment bestätigt, bei dem hitzeschockbehandelte und unbehandelte *Escherichia coli* Zellen verglichen wurden. Darüber hinaus wurden in den Einzelzelldaten im Vergleich zur Populations-Analyse Hinweise auf einzigartige Reaktionen bei der Synthese von Sekundärmetaboliten und der CRISPR-Cas-Editierung gefunden, was die Bedeutung der Untersuchung der Heterogenität seltener funktioneller Subpopulationen auf Einzelzellebene unterstreicht. Insgesamt wird erwartet, dass die verbesserten SCG- und SCT-Methoden, die in dieser Arbeit etabliert wurden, eine breitere Anwendung für ein besseres Verständnis der MDM-Diversität und -Funktion in der Umwelt ermöglichen.

Abstract

The vast majority of microbial life still remains undiscovered and understudied. We refer to these microorganisms as microbial dark matter (MDM) because they have not yet been successfully cultured. Within MDM hide potentially novel and important solutions for sustainable energy, bioremediation of contaminated environments, and the war against rising antibiotic resistance. The use of culture-independent methods to study microorganisms at the community-level, such as metagenomics and metatranscriptomics, have significantly advanced our understanding of MDM. However, these methods still struggle to reliably assemble individual genomes and transcriptomes, especially from low abundant organisms in highly diverse communities. Strain variation, the misattribution of sequences, highly repetitive sequence regions, and mobile genetic elements are a few of the problems that metagenomics faces. Likewise, in metatranscriptomics, the natural phenotypic heterogeneity of cells and diversity of microbial communities, results in complex transcriptional profiles that cannot be fully captured. Therefore, single-cell genomics (SCG) and transcriptomics (SCT), which together are referred to as single-cell ‘omics (SC ‘omics), were developed to overcome the disadvantages of metagenomics and metatranscriptomics by enabling the analysis of an individual cell.

The application of SCG has become an important tool for expanding our knowledge of MDM, for example, by enabling the recent discovery of several novel candidate phyla, which are currently only represented by single-amplified genomes (SAGs). However, complete SAGs from many organisms, especially minority members, are statistically hard to capture due to the high costs and many technical challenges throughout the workflows. On the other hand, microbial SCT is faced with the many challenges of working with RNA, such as the short half-life of mRNA and low levels of gene expressions, which is why SCT has not yet been widely applied in microbiology. Thus, the anticipated effects of SC ‘omics have not yet been fulfilled.

In a typical SCG workflow, after samples are collected, cells can be labeled with fluorescent dyes prior to single-cell isolation. After isolation, the cells are lysed and the genome has to be amplified for subsequent library preparation, which is followed by sequencing and data analysis. In this thesis, difficulties in the cell labelling, isolation, lysis, and whole genome amplification (WGA) steps were improved upon to overcome remaining challenges in SCG. First, a targeted-cell labeling approach was established, which enabled the enrichment of low abundant microorganisms from environmental samples. This approach aided in the discovery of novel phylogenies and metabolisms from rare members of the microbial community, which would have otherwise been overlooked by conventional metagenomics. Additionally, by targeting organisms of interest, this approach helped to reduce the costs of SCG by preventing the need to sequence tens of thousands of single-cells in order to access low abundant minority members. Next, improvements were made to the cell isolation and cell lysis steps to minimize both physical cell damage and DNA degradation, respectively, which helped to increase the success of the downstream genome amplification step. As for the WGA, a volume reduction approach was applied to significantly improve genome coverage uniformity and completeness. These findings highlighted the unnecessary need for further volume reduction down to nL or pL and costs could be reduced by 97.5%. It is anticipated that these advancements will increase the throughput of SCG and encourage the use of this approach in more research groups.

Since SCG alone only provides information on phylogeny, genetic structure and metabolic potentials, but not on the actual activity of a cell, a microbial SCT pipeline was developed in this thesis, to help understand the cell's individual functions in a community. Currently, very few methods for microbial SCT exist and the ones that do, remain widely unused outside of their respective groups due to their difficult handling and low accessibility. Therefore, in this study, modifications and improvements to a eukaryotic single-cell RNA sequencing (RNA-seq) method were made to enable its use in prokaryotes. Importantly, the addition of DTT in the lysis buffer was found to likely inhibit DNase I, leading to DNA contamination. The single-cell RNA-seq results herein revealed reliable transcriptional profiles when compared to bulk RNA-seq samples. This was also confirmed through a proof of principle experiment comparing heat-shock and non-treated *Escherichia coli* cells. Furthermore, evidence for unique responses involved in secondary metabolite synthesis and CRISPR-Cas editing were found upregulated in the single cell versus the bulk data, highlighting the importance for studying heterogeneity of functional subpopulations at the single-cell level. Overall, the improved SCG and SCT methods established in this work are anticipated to allow for more widespread use for further understanding of MDM diversity and function in the environment.

Scientific Publications

The following publications have been published or submitted for publication and include results reported in this dissertation:

1. Dam, H. T., Vollmers, J., **Sobol, M. S.**, Cabezas, A. & Kaster, A.-K. Targeted cell sorting combined with single cell genomics captures low abundant microbial dark matter with higher sensitivity than metagenomics. Front. Microbiol. 11, 1377 (2020).
2. Kaster, A. K. & **Sobol, M. S.** Microbial single-cell omics: the crux of the matter. Appl. Microbiol. Biotechnol. 104, 8209–8220 (2020).
3. **Sobol, M. S.** & Kaster, A.K. Gezielte Zellsortierung in der Einzelzellgenomik. BIOspektrum 2021 273 27, 274–276 (2021)**. ***
4. Zoheir, A.E., **Sobol, M.S.**, Ordonez, D., Kaster, A.K., Niemeyer, C.M., & Rabe, K.S. A three-colour biosensor reveals multimodal stress response at the single cell level and the spatiotemporal dynamics of biofilms. (*under review*).
5. **Sobol, M. S.** & Kaster, A.-K. Improving multiple displacement amplification for microbial single-cell genomics. (*under review*).

*denotes non-peer-reviewed publication

Abbreviations

ANI	Average nucleotide identity
CAG(s)	Co-assembled genome(s)
DMA	Droplet microarray
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
FACS	Fluorescence-activated cell sorting
fg	Femtogram
FISH	Fluorescence *in situ* hybridization
FSC	Forward scatter
GOLD	Genomes OnLine Database
IMG	Integrated Microbial Genomes database
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAG(s)	Metagenome-assembled genome(s)
MALBAC	Multiple Annealing and Looping Based Amplification Cycles
MDA	Multiple displacement amplification
MDM	Microbial dark matter
mL	Milliliter
NCBI	National Center for Biotechnology Information
ng	Nanogram
nL	Nano-liter
OD	Optical density
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
pg	Picogram
pL	Pico-liter
PTA	Primary Template-directed Amplification
RI	RNase Inhibitor
RNA	Ribonucleic acid

rRNA	Ribosomal RNA
RT	Reverse Transcription
SAG(s)	Single-cell amplified genome(s)
SCG	Single-cell genomics
SCT	Single-cell transcriptomics
SSC	Side scatter
UMI	Unique molecular identifier
WWTP	Wastewater treatment plant
WGA	Whole genome amplification
WTA	Whole transcriptome amplification

Table of Contents

Zusammenfassung-- iv

Abstract -- vii

Scientific Publications -- ix

Abbreviations--- x

Table of Contents -- xii

List of Figures --xvi

List of Tables --- xviii

1 Introduction ---1

1.1 Shedding light on the dark side of microbial life -- 1

1.2 New views on microbial diversity through culture-independent sequencing ------------------- 1

1.3 The limitations of meta ‘omic approaches -- 4

1.4 Microbial single-cell genomics: a new era -- 5

1.5 Single-cell genomics’ challenges and solutions --- 6

1.5.1 Sample collection ---7

1.5.2 Cell labeling --7

1.5.3 Single-cell isolation --8

1.5.4 Cell lysis ---9

1.5.5 Whole genome amplification (WGA) ---9

1.6 The next step: microbial single cell transcriptomics (SCT) -------------------------------- 14

1.6.1 Challenges in microbial SCT --16

1.6.2 Current methodology for microbial SCT --17

1.7 Summary and objectives -- 18

2 Materials and Methods -- 20

2.1 Chapter 1: Improving microbial single-cell genomics -- 20

2.1.1 Targeted-cell sorting with fluorescence in situ hybridization --------------------------------20

2.1.1.1 Sample collection and preparation ---------- 20
2.1.1.2 In solution, fixation-free fluorescence in situ hybridization ---------- 21
2.1.1.3 Targeted cell sorting of labeled cells ---------- 21
2.1.1.4 Multiple displacement amplification ---------- 22
2.1.1.5 16S rRNA gene amplification and screening ---------- 22
2.1.1.6 DNA extraction for metagenome sequencing ---------- 23
2.1.1.7 Library preparation for metagenome and single cell genome sequencing ---------- 23
2.1.1.8 Read processing and assembly ---------- 24
2.1.1.9 Genome assessment and co-assembly ---------- 24
2.1.1.10 Coverage assessment and binning ---------- 25
2.1.1.11 Phylogenetic analysis of Chloroflexi genomes ---------- 25
2.1.1.12 Genome analysis ---------- 26
2.1.1.13 Pyrosequencing ---------- 26
2.1.2 Improving cell isolation with a single-cell printer ---------- 27
2.1.2.1 Bacterial growth and FISH ---------- 27
2.1.2.2 Labeled cell-sorting ---------- 28
2.1.3 Improving cell lysis and multiple displacement amplification ---------- 28
2.1.3.1 Bacterial cell isolation ---------- 29
2.1.3.2 Cell lysis ---------- 29
2.1.3.3 Multiple displacement amplification ---------- 30
2.1.3.4 Library preparation and sequencing ---------- 30
2.1.3.5 Data processing and analysis ---------- 31
2.1.4 MDA test on a droplet microarray ---------- 32
2.1.4.1 Reducing evaporation issues with glycerol ---------- 32
2.1.4.2 MDA on the DMA ---------- 32
2.2 Chapter 2: Improving microbial single-cell transcriptomics ---------- 33
2.2.1 Improving reverse transcription for low cell inputs ---------- 33
2.2.1.1 Bacterial growth conditions and cell isolation ---------- 33
2.2.1.2 RNA extraction ---------- 34
2.2.1.3 Reverse transcription and PCR ---------- 34
2.2.2 Oxford Nanopore direct-cDNA sequencing ---------- 35
2.2.2.1 Bacterial growth and RNA extraction ---------- 35
2.2.2.2 Polyadenylation ---------- 35
2.2.2.3 Library preparation and RNA sequencing ---------- 36
2.2.2.4 Data processing and analysis ---------- 36

2.2.3 Modification of the REPLI-g single-cell WTA kit for prokaryotes -- 36
2.2.3.1 Bacterial growth conditions -- 37
2.2.3.2 Cell isolation and lysis -- 37
2.2.3.3 WTA and library preparation -- 37
2.2.3.4 Bulk RNA extraction and library preparation -- 39
2.2.3.5 Data Processing -- 40
2.2.3.6 Statistical Analysis -- 41
3 Results and Discussion -- 42
3.1 Chapter 1: Improving microbial single-cell genomics -- 42
3.1.1 A modified targeted-cell sorting method enriches minority taxa -- 42
3.1.2 Targeted-cell sorting captures rare taxa and their metabolic potential -- 46
3.1.3 Cell printers cause less damage to FISH-treated cells -- 49
3.1.4 Improved cell lysis provides better genome recovery -- 51
3.1.5 MDA volume reduction in multi-well plates improves genome coverage and uniformity ---------- 52
3.1.6 MDA volume reduction with DMA -- 60
3.1.7 Conclusion and outlook -- 63
3.2 Chapter 2: Improving microbial single-cell transcriptomics -- 65
3.2.1 Improving reverse-transcription for low cell inputs -- 65
3.2.2 Oxford Nanopore direct-cDNA -- 66
3.2.3 Optimizing single-cell whole-transcriptome amplification for prokaryotes -- 67
3.2.4 Modified single-cell WTA generates reliable scRNA-seq data -- 71
3.2.5 Heterogeneous single-cells still differentiate between treatment type -- 76
3.2.6 Microbial SCT reveals rare and unique activity -- 78
3.2.7 Conclusion and outlook -- 81
4 Future Work -- 84
5 References -- 85
6 Appendix -- 110
Appendix Table 1. MDA sample summary -- 110
Appendix Table 2. MDA read processing summary -- 111
Appendix Table 3. MDA read mapping summary -- 112
Appendix Table 4. MDA assembly statistic summary -- 113

Appendix Table 5. scWTA gene coverage summary -- 114

Appendix Table 6. DESeq2 result summary for top 8 transcribed heat-shock genes ----------------- 115

Appendix Table 7. DESeq2 result summary for type I CRISPR cas and enterobactin genes ---------- 116

Appendix Figure 1. Genome coverage by sequencing effort 1.25 µL MDA reaction volumes-------- 117

Appendix Figure 2. Read coverage and density plots for Liu et al (2019) scRNA-seq data------------ 117

Appendix Figure 3. Read coverage across the 16S rRNA operon for heat-shock samples ------------ 118

Appendix Figure 4. Read coverage across the 16S rRNA operon for non-treated samples----------- 118

Appendix Figure 5. PCA of the 300 most variant genes-- 119

7 Curriculum vitae --120

List of Figures

Figure 1.1. Proportions of Bacterial and Archaeal genomes, and their respective sources, in the Genomes OnLine Database (GOLD) 3

Figure 1.2. General overview of a single-cell genomics pipeline 6

Figure 1.3. Overview of WGA methods...11

Figure 1.4. The difference between bulk RNA-seq and single-cell RNA-seq...........................14

Figure 1.5. General overview of a single-cell transcriptomics pipeline..................................15

Figure 1.6. Timeline for single cell transcriptome developments over the years 16

Figure 2.1. RNA-seq data analysis workflow with Nextflow 40

Figure 3.1. Taxonomic overview of Uruguayan WWTP samples 43

Figure 3.2. FACS analysis of in solution, fixation-free FISH applied to a mock community......44

Figure 3.3. Epifluorescence microscopic and FACS of targeted Chloroflexi cells 45

Figure 3.4. Comparing different single-cell lysis conditions 52

Figure 3.5. General MDA statistics overview 53

Figure 3.6. Read processing statistics 54

Figure 3.7. Genome coverage and uniformity bias 56

Figure 3.8. Single-amplified (SAG) assembly statistics...59

Figure 3.9. MDA reaction kinetics for glycerol testing 60

Figure 3.10. Mimicking single-cell sorting onto the DMA 62

Figure 3.11. Comparison of reverse transcriptase on low input E. coli total RNA 66

Figure 3.12. ONT direct-cDNA library quality overview 67

Figure 3.13. REPLI-g Single Cell WTA method overview...68

Figure 3.14. Read coverage and density of smpB and ssrA genes 69

Figure 3.15. Transcript characteristic summary for heat-shock and non-treated samples 72

Figure 3.16. Gene length and GC content bias in RNA-seq..74

Figure 3.17. Pseudo-bulk versus bulk RNA-seq analysis 75

Figure 3.18. Highest transcribed heat-shock genes from single-cells 77

Figure 3.19. Bulk vs. scRNA-seq transcription of Enterobactin and Type I CRISPR Cas genes ..79

List of Tables

Table 1.1. Overview of microbial single-cell genome amplification methods 12

Table 1.2. Overview of microbial single-cell RNA-seq methods 18

Table 2.1. MDA reagents by reaction volume 30

Table 2.2. Modified REPLI-g WTA Single Cell reaction volumes and components 38

Table 3.1. Overview of Chloroflexi CAGs and SAGs collected with targeted SCG 47

Table 3.2. Viability after cell sorting with FACS and B.SIGHT™.. 50

1 Introduction

1.1 Shedding light on the dark side of microbial life

Microorganisms constitute an estimated ~4x10^{29} total number of cells on Earth (Bar-On et al., 2018) and can be found in almost every habitat. They harbor an enormous potential for biotechnological applications, such as novel natural product discovery, bioenergy production, and bioremediation of harmful anthropogenic-introduced substances (Abou Seeda et al., 2017; Katz & Baltz, 2016; Kumar & Kumar, 2017; Mullis et al., 2019; Stincone & Brandelli, 2020). Importantly, they also mediate the transformation of major elements such as carbon or nitrogen on a global scale (Falkowski et al., 2008). Despite their global quantity and importance, less than 1% of prokaryotes are estimated to have been cultured and therefore remain uncharacterized, obscuring our knowledge of microbial diversity, metabolism, (eco)physiology, inter-organism interactions, and adaptive evolution (Hug et al., 2016; Lloyd et al., 2018; McDonald et al., 2012; Wu et al., 2009). We refer to these unknown microbes as “Microbial Dark Matter” (MDM). MDM likely remains uncultured due to their specific environmental and ecological needs that cannot be replicated easily in the lab (Stewart, 2012). New cultivation methods have been developed to help tackle this problem, but their ability to uncover large amounts of novel species is still lacking and they remain largely dependent on genomic data (Lewis et al., 2020; Wiegand et al., 2020). Further understanding of microbial diversity, function, and evolution requires cultivation-independent methods to uncover the remaining 99% of microbial species waiting to be characterized and be potentially used in biotechnological applications.

1.2 New views on microbial diversity through culture-independent sequencing

In 1977, through amplicon sequencing of the 16S rRNA gene, Carl Woese and George Fox fundamentally changed our view on the tree of life by greatly expanding our understanding on microbial diversity and dividing life into the three domains we know of today (Woese & Fox, 1977). Today, the 16S rRNA gene still remains the key gene for the identification and classification of prokaryotes. Amplicon sequencing brought forth a new era of culture-independent research, and it was realized that the discovery of novel organisms in the environment was far exceeding the number of cultured isolates. Now, there was a need to access more than just the 16S rRNA gene from these uncultured organisms in order to learn more about their physiology.

In 1996, Stein et al. developed the first approach to capture large genome fragments of microorganisms from an environmental sample (Stein et al., 1996), which was later termed "metagenomics" (Handelsman et al., 1998). This early approach was based on cloning, a process that is not only time and resource consuming but also prone to bias (Huber et al., 2009). Additionally, high sequencing costs limited the sequencing depth and therefore resolution of early metagenomic analyses. Since then, sequencing as well as metagenomic methods have greatly improved, significantly decreasing the cost and effort to the point that it is now feasible to obtain the "collective" metagenome from a complex environmental sample. Using a computational method called "binning", scientists can reconstruct individual genomes of different taxa from these metagenomes, i.e. metagenome-assembled genomes (MAGs) (Tyson et al., 2004). This genome-resolved metagenomics has further transformed the tree of life (Hug et al., 2016), enabled the discovery of many new uncultivated phyla (Anantharaman et al., 2016; Brown et al., 2015) (**Figure 1.1**), and extended our knowledge on the metabolic potential of many different bacterial and archaeal lineages (De Anda et al., 2021; Delmont et al., 2018; Murphy et al., 2021; Wiegand et al., 2020). However, metagenomics can only provide us with the overall genomic potential of a community, not which genes are currently expressed under certain environmental conditions, or even which organisms are currently active or dormant. These limitations still restrict the insights into the ecological role of novel, uncultured organisms as well as the metabolic function of novel genes found in their respective reconstructed genomes. Therefore, a complementary approach called metatranscriptomics was developed, which analyzes all expressed mRNA transcripts within a microbial community. When combined with metagenomics, this analysis enables researchers to quantify and compare the level of gene expression to understand more about the ecological function of specific organisms and/or genes within the community as well as how the overall community adapts to certain conditions (Desai et al., 2010). Other approaches that complement metagenomics and metatranscriptomics include metaproteomics and metametabolomics, each providing different insights into the function and activity of microbial communities through the analysis of proteins and metabolites, respectively (Beale et al., 2016). All together, these community-wide approaches are referred to as meta ‘omics analyses.

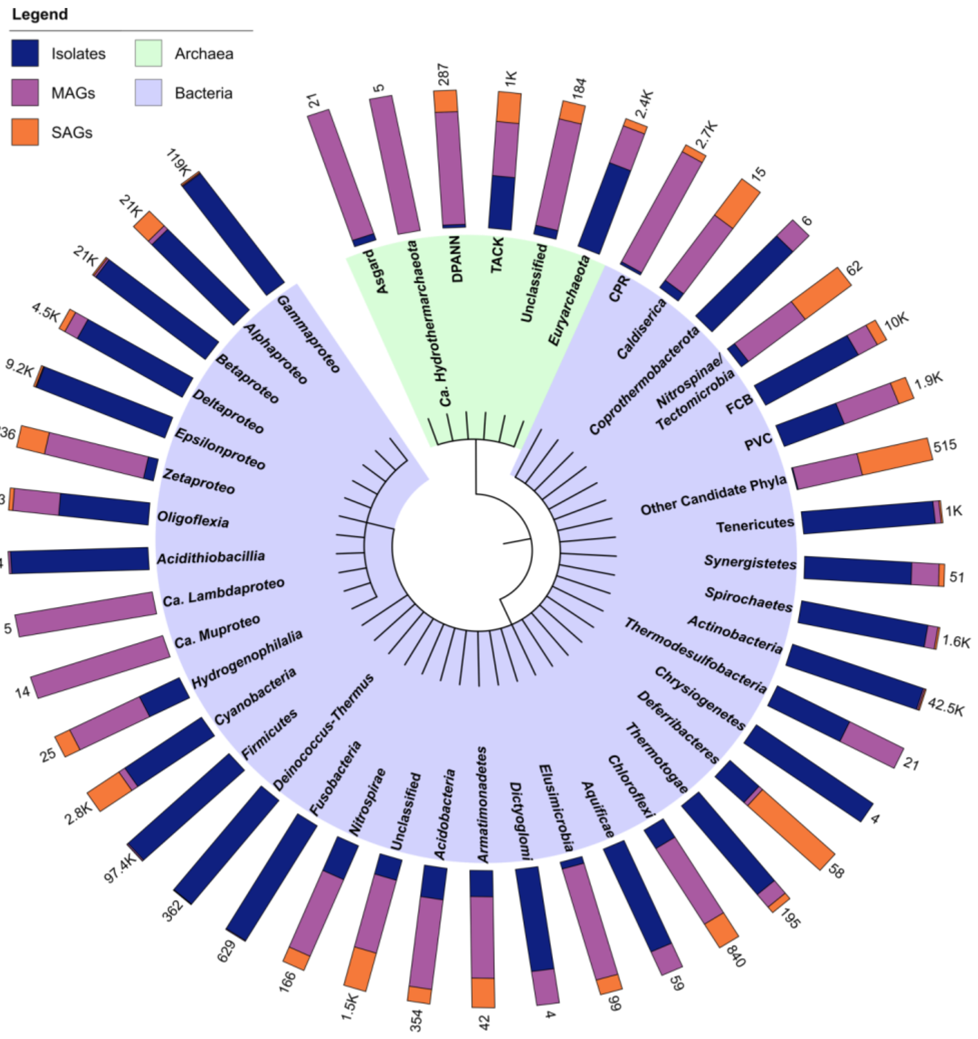


Figure 1.1. Proportions of Bacterial and Archaeal genomes, and their respective sources, in the Genomes OnLine Database (GOLD)

Cladogram of prokaryotes (Bacteria and Archaea) showing the relative proportions of isolate genomes, single-amplified genomes (SAGs), and metagenome-assembled genomes (MAGs) that make up the total number of genomes in each phylum. The taxonomy is based on National Center for Biotechnology Information (NCBI) (Sayers et al. 2020). Total genome numbers for each phylum are shown at the top of each bar. Data extracted from the Genomes OnLine Database (GOLD) in July 2020 (Mukherjee et al. 2019). Cladogram created with Interactive Tree of Life (iTOL) version 5 (Letunic and Bork 2019). -proteo -proteobacteria. Asgard Lokiarchaeota-Thorarchaeota-Odinarchaeota-Heimdallarchaeota. DPANN Diapherotrites-Parvarchaeota-Aenigmarchaeota-Nanoarchaeota-Nanohaloarchaeota. TACK Thaumarchaeota-Aigarchaeota-Crenarchaeota-Korarchaeota. FCB Fibrobacteres-Chlorobi-Bacteroidetes. PVC Planctomycetes-Verrucomicrobia-Chlamydiae. CPR Candidate Phyla Radiation. Published in Kaster and Sobol (2020).

1.3 The limitations of meta ‘omic approaches

Unfortunately, meta ’omics alone is still limited when applied to complex and/or highly heterogenous microbial communities. Heterogeneity is a common characteristic of microorganisms to adapt to environments with constant and rapid changes (González-Cabaleiro et al., 2017; Martins & Locke, 2015; Morawska et al., 2022). This observed heterogeneity between closely related organisms is due to bet-hedging, where random subpopulations of cells diversify their phenotypes as a risk-mitigation strategy (Ackermann, 2015; Morawska et al., 2022). Examples of bet-hedging strategies, reviewed in Morawska et al. (2022), include prokaryotic cell specialization for biofilm formation and quorum signaling, specialized persister and/or sporulated cells, and cell variants which use different nutrient resources. This phenotypic diversity leads to different expression patterns for otherwise identical cells, but can also take the form of actual genomic differences between members of the same species (so called “strain variations”). Strain variations especially complicate the binning and reconstruction of individual genomes of different species within a community (Dick et al., 2009). This effect can be most severe for low abundant organisms, since the quality of genome reconstruction is largely dependent on sequence coverage for assembly as well as coverage covariance based binning (Albertsen et al., 2013; Dam et al., 2020; Vollmers et al., 2017). MAGs are therefore often consensus genomes of all possible strain variants from one sample (Van Rossum et al., 2020). Another problem is the potential to misattribute contigs to the wrong genomes resulting in chimeric genomes not representing actual organisms and subsequent database error propagation as MAGs contaminated with more than one species are often uploaded in databases. Recently, it was shown that some publicly available MAGs were found to be as high as 48% contaminated within databases (Vollmers et al., 2022). Furthermore, highly repetitive sequences like those found in CRISPR regions (Acuña-Amador et al., 2018; Skennerton et al., 2013) are often not accurately assembled and 16S rRNA sequences as well as mobile genetic elements such as plasmids can often not be attributed to their host organisms (Dam et al., 2020; Maguire et al., 2020). As a result, insights into evolutionary mechanisms, like horizontal gene transfer, are lost.

Considering that metatranscriptomics gives us insight into the global expression profile of a community, it too struggles to resolve the heterogeneity of microbial gene transcription, which leads to the poor understanding of rare, functioning subpopulations of microorganisms (Bossert et al., 2018; Imdahl & Saliba, 2020; Kaster & Sobol, 2020; Picelli, 2017; Roberfroid et al., 2016). This is again especially the case for minority taxa, and/or low abundant transcripts. Consequently, ambiguous information about the organization and activity of genes within genomes makes it difficult to predict functional genes, metabolic pathways, and potential benefits of uncultured microbial species. The limitations of meta 'omics have spurred a new era of technological advancements, namely microbial single-cell 'omics, to complement these approaches.

1.4 Microbial single-cell genomics: a new era

Single cell genomics (SCG) enables the study of a single microbial cells' DNA and was developed to overcome the limitations of metagenomics. Since 2005, SCG has become a powerful tool for studying uncultivable organisms and delineating complex populations (Raghunathan et al., 2005). An increasing number of SAGs are available from public databases such as the National Center for Biotechnology Information (NCBI) GenBank (Sayers et al., 2020), and/or the Joint Genome Institute Genomes OnLine Database (GOLD) (Mukherjee et al., 2019), which includes all data from Integrated Microbial Genomes (IMG). As of July 2022, over 10,000 SAG sequencing projects have been deposited in GOLD (Mukherjee et al., 2019), of which many are classified as uncultured and potentially novel taxonomic groups (Becraft et al., 2016; Hedlund et al., 2014; Landry et al., 2017; León-Zayas et al., 2017; McLean et al., 2013; Swan et al., 2011) (**Figure 1.1**). Recently, a new reference database containing over 12,000 SAGs from the euphotic ocean was published (Pachiadaki et al., 2019), greatly expanding our knowledge on the diversity and complexity of marine microorganisms. However, even *via* single-cell 'omics, many species continue to elude analysis attempts, i.e. minority members in a microbial community and/or anaerobic organisms, as the current standard SCG workflow still has several drawbacks.

1.5 Single-cell genomics’ challenges and solutions

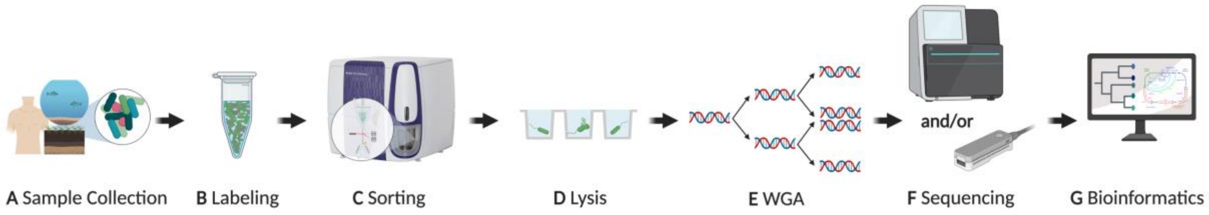


Figure 1.2. General overview of a single-cell genomics pipeline
A Unless analyzed immediately, environmental samples require deep-freezing in the presence of a cryoprotectant that preserves the integrity of the cell. **B** Cells are stained with a fluorescent dye, such as DAPI or SYBR® Green, however, they can also be specifically labelled. **C** Physical isolation of a single-cell can be performed by Fluorescent Activated Cell Sorting (FACS), cell printing (not shown), or microfluidics (not shown) into multi-well plates or other platforms. **D** After separation, the single cells are lysed to release their DNA. Today, most cell lysis in SC omics relies on an alkaline solution. **E** Since a typical prokaryotic cell only contains a few fg grams of DNA, multiple displacement amplification (MDA) can be used for whole genome amplification. **F** After library preparation, Next Generation Sequencing technologies like Illumina, Oxford Nanopore or PacBio (not shown) are available for sequencing. **G** After quality assessment, trimming, and/or normalization of the sequencing reads, bioinformatics tools can conduct the assembly, classification, ORF calling, and annotation of the genes. Created with BioRender.com. Modified from Kaster & Sobol (2020).

In general, the SCG workflow involves (A) sampling and preservation, (B) non-specific staining of microbial populations, (C) cell sorting, (D) cell lysis, (E) whole genome amplification (WGA), and (F) sequencing and (E) analysis (**Figure 1.2**) (Kaster & Sobol, 2020; Rinke et al., 2014). Many technical issues arise throughout the SCG pipeline due to difficulties in cell labelling (Müller & Nebe-Von-Caron, 2010) and premature cell lysis during sorting (e.g. anaerobic cells subjected to oxygen), leading to early DNA degradation prior to the downstream processes (Bellais et al., 2022). Furthermore, genome amplification bias can severely limit the completeness of the recovered SAGs, which in practice, varies widely from less than one percent to a complete finished genome (Clingenpeel et al., 2014; Stepanauskas et al., 2017). Primarily these issues have much to do with the fact that microorganisms are vastly different regarding their shapes, sizes, cell wall types, and cell abundances, making it difficult to apply one approach to a diverse sample (Kaster & Sobol, 2020). Additionally, due to the femtogram (fg) levels of DNA per a single cell, WGA often struggles to capture the entire genome, and is easily contaminated if the correct precautions are not taken. How these challenges affect the individual steps of the SCG workflow and what solutions can be applied, are discussed in detail below.

1.5.1 Sample collection

Most of the times, samples (especially from environmental samples) cannot be processed through the SCG workflow immediately after collection. Therefore, storage solutions that preserve the integrity of the cells is crucial for subsequent labeling and sorting steps. The current standard approach recommends flash-freezing the samples with liquid nitrogen in the presence of a cryoprotectant, such as glycerol or betaine (Rinke et al., 2014). Fixatives such as paraformaldehyde and ethanol may negatively impact downstream analysis (Clingenpeel et al., 2014). Working with sediment or soil samples adds a level of complexity since many cells will be attached to particles and/or aggregated in biofilms. Therefore, additional vortexing and centrifugation steps have to applied, but must be done carefully in order to not destroy the cells.

1.5.2 Cell labeling

Conventional SCG uses nucleic acid stains like SYBR Green I for non-specific staining of microbial cells prior to cell sorting (**Figure 1.2B**). However, this method is inefficient when specific members of a community are to be targeted, especially minority taxa, since they are statistically harder to sort. Thus, this approach becomes very costly when analyzing highly diverse microbial communities (Dam et al., 2020; Kaster & Sobol, 2020). In contrast, taxon-specific or function-based labeling and sorting of targeted cells facilitates enrichment of specific taxa of interest (Dam et al., 2020; Doud et al., 2019; Hatzenpichler et al., 2016; Pratscher et al., 2018).

One approach for targeted sorting utilizes fluorescence *in situ* hybridization (FISH), a method which employs fluorescently-labeled oligonucleotides to target ribosomal RNA (rRNA) within a cell (Pernthaler et al., 2001), in conjunction with single-cell sorting to isolate certain microorganisms (Haroon et al., 2013; Podar et al., 2007; Yilmaz et al., 2010). In traditional FISH protocols, cells are chemically fixed onto glass slides and permeabilized with paraformaldehyde, but in order to be compatible with FACS and downstream processes, the previously mentioned methods removed the fixation and permeabilization steps (Haroon et al., 2013; Podar et al., 2007;

Yilmaz et al., 2010) as these treatments have been shown to negatively impact the downstream processes and subsequent genome recovery (Clingenpeel et al., 2014). Thus, a targeted approach that does not diminish downstream genome amplification and is capable of labeling cells within complex microbial communities, especially when interested in minority members, is needed.

1.5.3 Single-cell isolation

Several different methods can be used for single cell isolation, such as microfluidics, micromanipulation, and fluorescence-activated cell sorting (FACS) (**Figure 1.2C**). Microfluidic- and optofluidic-based systems (Gole et al., 2013; Lan et al., 2017), as well as micromanipulation (Grindberg et al., 2011; Woyke et al., 2010), have the advantage of sorting cells based on their morphology and applying less physical stress to the cell. Some setups even allow for cell separation, lysis, and amplification performed in one closed system at nL or even pL volumes (Blainey et al., 2011; Landry et al., 2017; Marcy, Ishoey, et al., 2007; Marcy, Ouverney, et al., 2007; Xu et al., 2016), however, these devices remain limited in cell throughput, accessibility to experimental set ups, and the successful recovery of non-contaminated amplified products (Kaster & Sobol, 2020). Thus, FACS has become the most commonly used method for single cell isolation due to its high throughput, flexibility with the use of different fluorescence signals, and the fact that it is commercially available (Rinke et al., 2014; Stepanauskas & Sieracki, 2007; Woyke et al., 2017). However, the main limitations of FACS include the inability to microscopically examine cells, further miniaturization of downstream reaction volumes, difficulty in sorting under anoxic conditions, and the strong physical stress it applies which can prematurely cause cell lysis during sorting (Blainey, 2013; Mollet et al., 2008; Wiegand et al., 2021). Premature cell lysis is especially an issue for downstream steps in the SC 'omics pipeline and can occur during sorting due to damaged cell walls caused by sample preparation (e.g. FISH treatment, as discussed above). If a cell is lysed prior to the cell lysis step (**Figure 1.2D**), the DNA is likely to be lost during sorting and can cause contamination when sorted within the droplets of other cells (Wiegand et al., 2021). Or, the DNA may still be contained in the droplet of the original cell, but will be fragmented and/or degraded when lysis buffers are applied, which leads to lower or no genome recovery.

To help mitigate problems with downstream applications caused by FACS, a cell printing technology was recently developed (Gross et al., 2013; Riba et al., 2016). This technology uses modified inkjet printer heads to more gently deposit cells in smaller volumes than FACS and selects cells based on their morphology with bright-field imaging. Furthermore, because of its small size and low sorting buffer requirements, it can be easily placed inside anaerobic tents and there is no need to remove oxygen from tens of liter of sheath fluid (compared to FACS), enabling more accessible sorting of anaerobic microorganisms. Establishing this technology for microbial SCG will enable greater success of single amplified genomes (SAGs) and more reliable detection of potential contamination (Wiegand et al., 2021).

1.5.4 Cell lysis

As briefly mentioned above, cell lysis efficiency plays a critical role in the success rate of SC ‘omics but is challenging due to the natural diversity of microbial cell walls which cause cells to either be prematurely lysed during cell sorting (i.e. gram negative cells) or too difficult to lyse (i.e. gram positive cells) (Liu et al., 2018; Rinke et al., 2014; Stepanauskas, 2012). To be effective, cell lysis need to accomplish releasing nucleic acids from the cell without damaging them and must not interfere with downstream reactions (Bäumer et al., 2018; Clingenpeel et al., 2014). Lysis which strongly fragments DNA into short fragments will not be suitable for amplification steps since standard WGA polymerases require larger templates for synthesis (Blainey, 2013). Currently, alkaline lysis is the most widely used method; however, more efficient lysis of cells from complex communities or cells with tougher cell walls may be accomplished by using a combination of freeze-thaw cycles, chemical, or enzymatic lysis methods (Hall et al., 2013; He et al., 2016; Liu et al., 2018; Stepanauskas et al., 2017). Due to the importance of cell lysis for successful single-cell sequencing, one should consider the type of sample or taxa of interest for the best results.

1.5.5 Whole genome amplification (WGA)

The WGA step is crucial for generating a sufficient amount of input DNA for library preparation and subsequent sequencing, as a typical microbial cell only contains a few femtograms (fg) of DNA (Hedlund et al., 2014; Rodrigue et al., 2009). Several different WGA methods have been developed and improved upon over the years. These methods can be categorized as Polymerase Chain Reaction (PCR)-based amplification, isothermal amplification, and hybrid, which combines both methods (Gawad et al., 2016). Pure PCR-based methods, such as degenerate oligonucleotide primed PCR (DOP-PCR) (Telenius et al., 1992), were not successfully applied to microbial single-cells, likely because of sensitivity issues. The first method to amplify DNA from a single bacterial cell was the so called multiple displacement amplification (MDA) (Raghunathan et al., 2005) (**Figure 1.3A; Table 1.1**). MDA is an isothermal method that uses a phi29 polymerase, which has a lower error rate (1 in 10^6 bases) compared to standard polymerases used in PCR, high fidelity for the template, 3′ → 5′ exonuclease proofreading activity, and generates fragments larger than 10 kb (Dean et al., 2001; Esteban et al., 1993; Paez et al., 2004; D. Y. Zhang et al., 2001) (**Table 1.1**). Currently, MDA remains one of the most widely applied methods for amplifying DNA from microbial single cells for these reasons (Kaster & Sobol, 2020).

Unfortunately, MDA also constitutes one of the major limitations in single cell sequencing due to its high costs (**Table 1.1**), as well as it’s bias against high GC regions, which leads to uneven genome amplification (Lasken, 2009; Lasken & Stockwell, 2007; Sabina & Leamon, 2015). Furthermore, artifacts like chimeras and non-specific products can be produced. These artifacts are thought to occur randomly since sequences that are over-represented in one MDA reaction can be under-represented in another (Lasken & Stockwell, 2007; Sabina & Leamon, 2015). However, some have found these effects to be reproducible due to the fact that decreased template copy number increases bias and that certain sequences are simply not amplified at all (Dean et al., 2001; Lasken, 2009; Lasken & Stockwell, 2007; Wu et al., 2006). As a result, treatments such as post-amplification endonuclease and post-amplification normalization by nuclease degradation of dsDNA have been used to reduce chimeric sequences (Zhang et al., 2006) and highly abundant sequences (Rodrigue et al., 2009), respectively.

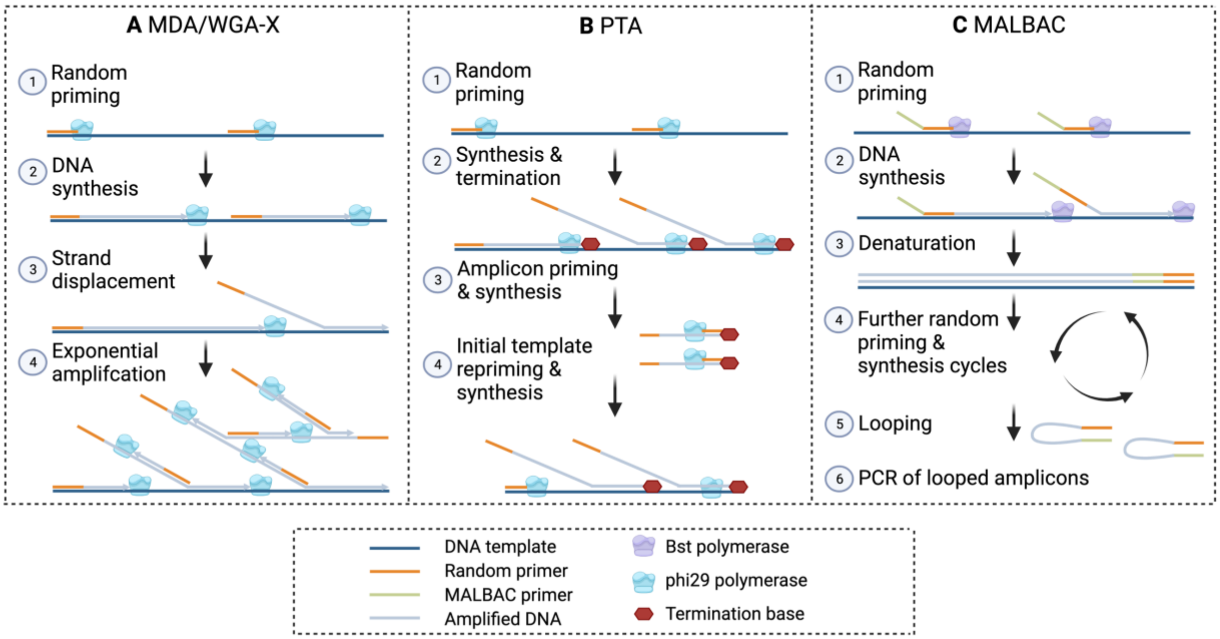


Figure 1.3. Overview of WGA methods
A MDA, Multiple Displacement Amplification; WGA-X, Whole Genome Amplification – X. **B** PTA, Primary Template-directed Amplification. **C** MALBAC, Multiple Annealing and Looping Based Amplification Cycles. Made with Biorender.com.

Other approaches have worked to improve MDA its self, such as WGA-X™, which uses a more thermostable phi29 polymerase for better amplification of high GC organisms (Stepanauskas et al., 2017) (**Figure 1.3A; Table 1.1**). However, lower genome coverage for low GC organisms compared to standard MDA is reported. More recently, Primary Template-directed Amplification (PTA) was developed, which employs exonuclease-resistant terminators to create smaller amplicons that undergo limited subsequent amplification to limit overrepresentation of random positions and reduce error propagation (Gonzalez-Pena et al., 2021) (**Figure 1.3B; Table 1.1**). While this method looks promising to reduce amplification bias, the approach is still in the alpha testing stage for microorganisms (https://www.bioskryb.com/resolvedna-microbiome-alpha/) and quite expensive. The hybrid method, Multiple Annealing and Looping Based Amplification Cycles (MALBAC), combines PCR and MDA methods to successfully reduce amplification bias (Lu et al., 2012; Zong et al., 2012) (**Figure 1.3C; Table 1.1**). Yet, MALBAC remains widely unused in microbial SCG, because the Bst and Taq polymerases have higher error rates and lack proof-reading capability (De Bourcy et al., 2014). Thus, further work needs to be done to optimize MALBAC, possibly with phi29 or less error-prone enzymes (Lasken, 2013).

Table 1.1. Overview of microbial single-cell genome amplification methods
MALBAC, Multiple Annealing and Looping Based Amplification Cycles; MDA, Multiple Displacement Amplification; WGA-X, Whole Genome Amplification – X; PTA, Primary Template-directed Amplification.

Method Characteristics	**Classic MDA**[1-3]	**MDA *via* WGA-X™**[2]	**MDA *via* PTA**[3,4]	**MALBAC**[5]
Specific Primers	no	no	no	yes
Enzyme Type	phi29	EquiPhi29™	phi29	Bst & Taq polymerase
Proof-reading	yes	yes	yes	no
Strand Displacement	yes	yes	yes	yes
Amplification Type	Exponential	Exponential	Quasi-linear	Quasi-linear
Product Length (nt)	>10,000	>10,000	250-1,500	500-1,500
Average Genome Coverage for *E. coli*	~10 to 80%	36 ± 21%	≥92%	~80%
Recommended Reaction Volume*	50 µL	10 µL	20 µL	65 µL
Approx. Costs per 1.0 µL Reaction	0.48 $	0.14 $	1.50 $	0.72 $

* Reaction volumes include sorting, lysis and neutralization buffer volumes as well as WGA reagents and/or fluorescent dyes to monitor the reaction recommended by the manufacturer or authors of the study. [1]Marcy et al. (2007) [2] Stepanauskas et al. (2017), [3] BioSkyrb Genomics, Inc., [4] Gonzalez-Pena et al. (2021), [5] De Bourcy et al. (2014).

Even though there is hope to reduce amplification bias in microbial WGA, statistically, inconsistency and bias between the DNA amplification of millions of templates will still persist (C.-Z. Zhang et al., 2016). In addition, WGA methods are highly sensitive to contamination due to the low amounts of DNA from a single-cell. Prior decontamination of reagents with UV (Woyke et al., 2011) can help to remove common reagent contaminants, but this does not prevent other sources of endogenous and/or exogenous contaminants, which become more amplified in larger WGA reaction volumes due to reduced polymerase specificity (Hutchison et al., 2005). Therefore, through bioinformatics, contamination in SAGs needs to be analyzed and removed prior to downstream analysis. Moreover, the large recommended reaction volumes of these WGA methods also quickly become very costly when applied to hundreds of single-cells (**Table 1.1**).

These high costs limit the depth that samples can be analyzed, preventing, for example, minority taxa from being captured with SCG.

Therefore, a methodically simpler solution is to reduce WGA's reaction volume. Reduction of total WGA volume has been shown to increase the concentration of the template and lessens the chance for background contamination to be amplified (Hutchison et al., 2005). Furthermore, this approach also significantly reduces the high costs of WGA (**Table 1.1**). Previous studies have applied this approach at sub-nL and pL volumes in microfluidic devices (Blainey et al., 2011; De Bourcy et al., 2014; Marcy, Ishoey, et al., 2007; Nishikawa et al., 2015; Rhee et al., 2016; Ruan et al., 2020; Sidore et al., 2015), nanowells (Goldstein et al., 2017; Gole et al., 2013), planar surfaces (Leung et al., 2016; Rezaei et al., 2021), and hydrogels (Xu et al., 2016). However, these approaches and their devices remain largely unused outside of their respective publications, likely because most microfluidic chips and other platforms are not commercially available and therefore hard to access and implement in other research groups. Also, many lack the throughput needed for microbial SCG and/or they sort based on Poisson distributions of cells resulting in high unoccupancy and cell loss (Collins et al., 2015), which is not applicable for studies analyzing rare populations. Hence, the establishment of a reliable and easy-to-use volume reduction method is needed to widen the accessibility and application of microbial SCG.

1.6 The next step: microbial single cell transcriptomics (SCT)

Most natural microbial communities are complex and made up of phenotypically diverse organisms that are divided into clonal sub-populations. Clonal populations are often considered functionally uniform, yet studies show that even clonal populations can be further divided into phenotypically heterogeneous sub-populations (Ackermann, 2015; Martins & Locke, 2015). Much of cellular heterogeneity in isogenic populations has so far been contributed to the so-called "bet-hedging" strategy, where a population increases its chance for survival under changing environmental conditions (e.g. anoxia, antibiotic stress, starvation) by dividing important metabolic and regulatory functions amongst individuals (Morawska et al., 2022). Additionally, this approach is also used to divide labor amongst cells for more efficient growth and/or biofilm formation (Morawska et al., 2022). While much of the work on elucidating cell-to-cell heterogeneity has been discovered with fluorescence microscopy, microscopy-based methods are not high-throughput and generally require knowledge on the communities function (Brennan & Rosenthal, 2021).

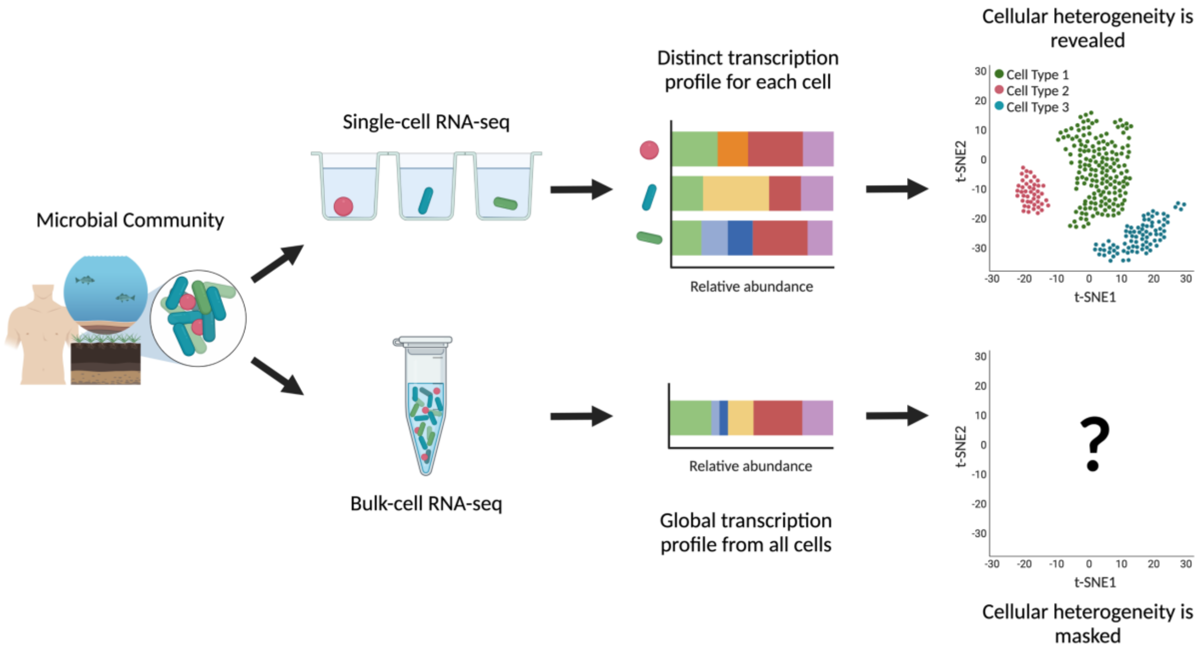

Cellular heterogeneity is revealed
Distinct transcription profile for each cell
Single-cell RNA-seq
Microbial Community
Cell Type 1
Cell Type 2
Cell Type 3
t-SNE2
t-SNE1
30
20
10
0
-10
-20
-30
Relative abundance
Bulk-cell RNA-seq
Relative abundance
Global transcription profile from all cells
?
Cellular heterogeneity is masked

Figure 1.4. The difference between bulk RNA-seq and single-cell RNA-seq
Microbial scRNA-seq analysis reveals cellular heterogeneity that would otherwise be masked by bulk RNA-seq methods. Furthermore, low abundant transcripts may not be captured. Adapted from 10xgenomics.com, created with Biorender.com.

Furthermore, bulk RNA-sequencing (RNA-seq) is not helpful as it only provides the universal transcriptional profile of an entire community, thus any chance to analyze cellular heterogeneity is masked (**Figure 1.4**). Hence, the necessity of studying transcription at single-cell resolution becomes important in order to better understand the forces driving and organizing cellular heterogeneity and the ecological effects of phenotypically different subpopulations (de Jager & Siezen, 2011; Kanter & Kalisky, 2015) (**Figure 1.4**).

The single-cell RNA-seq (scRNA-seq) workflow is similar to SCG, but with some additional steps (**Figure 1.5**). After cells are individually isolated and lysed (**Figure 1.5A-D**), the RNA must first be converted to cDNA (**Figure 1.5E**) prior to whole transcriptome amplification (WTA) (**Figure 1.5F**)**.** However, compared to genomic analysis, WTA analysis for microorganisms at the single-cell level is even more challenging and most methods for eukaryotic scRNA-seq are not applicable to microorganisms.

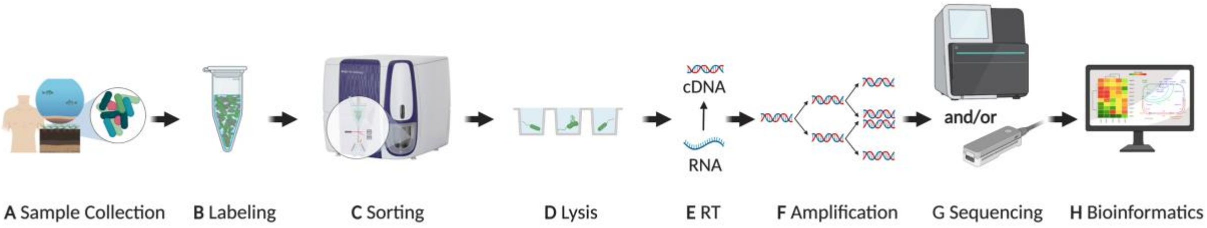


Figure 1.5. General overview of a single-cell transcriptomics pipeline
A Unless analyzed immediately, samples require deep-freezing in the presence of a cryoprotectant that preserves the integrity of the cell. **B** Cells can be stained with a fluorescent dye, such as DAPI or SYBR® Green, but precaution has to be taken to not alter the transcriptome of the cells. **C** Physical isolation of a single-cell can be performed by Fluorescent Activated Cell Sorting (FACS), cell printing, or microfluidics (not shown) into multi-well plates or other platforms. **D** After separation, the single cells are lysed to release their RNA without degrading it. **E** RNA must first be converted to double-stranded cDNA *via* reverse transcription prior to amplification. **F** Since a typical prokaryotic cell only contains a few fg grams of RNA, multiple displacement amplification (MDA) can be used for whole transcriptome amplification. **G** After library preparation, Next Generation Sequencing technologies like Illumina, Oxford Nanopore or PacBio (not shown) are available for sequencing. **H** After quality assessment, trimming, and/or normalization of the sequencing reads, bioinformatics tools can conduct the transcriptome assembly, gene annotation, gene counting, and differential expression analysis. Created with BioRender.com. Modified from Kaster & Sobol (2020).

1.6.1 Challenges in microbial SCT

SCT of Eukaryotes has been widely applied (Adil et al., 2021; Wolfien et al., 2021) (**Figure 1.6**), but most of these methods are not suitable for microbial SCT due to the vast differences between a eukaryotic and microbial cell. Thus, comparatively little progress has been made in the field of microbial SCT when compared to eukaryotic scRNA-seq (**Figure 1.6**). A single eukaryotic cell contains approximately two orders of magnitude more total RNA per a cell than prokaryotic cells (Brennan & Rosenthal, 2021; Imdahl & Saliba, 2020) and only a small is mRNA, as rRNA and tRNA molecules usually represent over 90% of the total RNA. Because most eukaryotic SCT methods select for polyadenylated mRNA transcripts with oligo (dT) primers, the lack of a polyA tail on most mRNAs in Prokaryotes means that the large fraction of rRNA and tRNA cannot be depleted using these methods. In addition, microbial mRNA has an average half-life of 10 minutes compared to 10 hours for eukaryotic mRNA, which means methods need to be adapted to process the RNA in a shorter period of time (Brennan & Rosenthal, 2021). Lastly, like SCG, the same challenge with cell lysis and differences in cell wall structure also explains why not all methods for eukaryotic SCT can be applied to microorganisms (Brennan & Rosenthal, 2021; Kaster & Sobol, 2020; Zhang et al., 2018).

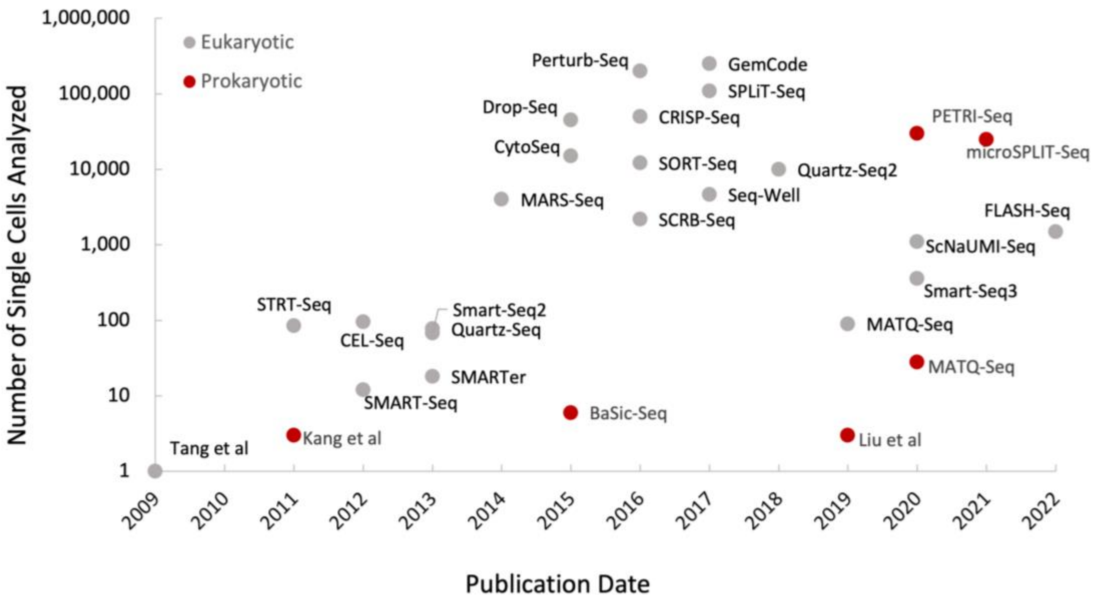


Figure 1.6. Timeline for single cell transcriptome developments over the years
Timeline of notable developments for both eukaryotic (grey) and prokaryotic (red) single cell transcriptomics. Graph modified from Blattman et al. (2020).

1.6.2 Current methodology for microbial SCT

Even with this long list of challenges, fortunately several new approaches were developed in the last decade for microbial SCT analysis (Blattman et al., 2020; Imdahl et al., 2020; Y. Kang et al., 2015, 2011; Kuchina et al., 2021; Liu et al., 2019; J. Wang et al., 2015) (**Figure 1.6**). To overcome the lack of polyA tails on microbial mRNA, these methods use random primers to non-specifically prime all RNA. All methods largely differ in their cell isolation strategies, amplification approach, number of cells they can process in a given experiment, and the amount of mRNA they capture (**Table 1.2**). The method by Kang et al. (2011,2015) was not included in **Table 1.2** because only microarray analysis was used and coverage was therefore not comparable. In general, these studies already provide the basic framework for microbial SCT, however, they have so far only been carried out on model organisms and many technical challenges still remain due to difficulty in accessing specialized equipment (i.e. microfluidic chips, micropipette) (Liu et al., 2019; Wang et al., 2015), biased amplification (Chen et al., 2017; Picelli, 2017; Zhang et al., 2018), and contamination (Wang et al., 2015). Additionally, previous studies have reported detecting thousands of transcripts from a single-cell (Liu et al., 2019; Wang et al., 2015), whereas more recent studies only report a few hundred (**Table 1.2**). The most likely explanation is not due to an insensitivity with newer methods, but possibly caused by DNA contamination which produced false-positive transcripts. This is supported by the fact that on average, transcripts are present in less than one copy per a gene (Imdahl & Saliba, 2020), as well as the fact that we found evidence for DNA contamination in the scRNA-seq data from Liu et al. (2019). Therefore, the transcript coverage of only a few hundred genes is more realistic. However, in order to expand the use microbial SCT, improvements to cell isolation, cell lysis, DNA removal, and the amplification strategies of these existing methods are still needed.

Table 1.2. Overview of microbial single-cell RNA-seq methods
SPIA, Single Primer Isothermal Amplification; MDA, Multiple Displacement Amplification; PCR, Polymerase Chain Reaction, FACS, Fluorescence-activated Cell Sorting.

WTA Method	**BaSic-seq[1]**	**REPLI-g WTA[2]**	**MATQ-seq[3]**	**PETRI-seq[4]**	**microSPLiT[5]**
Organism	*Synechocystis* sp. PCC 6803	*Porphyromonas somerae*	*Salmonella enterica*	*Escherichia coli*, *Staphylococcus aureus*	*Escherichia coli*, *Bacillus subtilis*
Cell Isolation	micropipette	microfluidic	FACS	N/A	N/A
Amplification Strategy	SPIA	MDA	PCR	PCR	PCR
No. of Single Cells/Experiment	6	10	~25	~30,000	~25,000
Gene Coverage/ Single cell	34 - 99%	~75%	~5%	2-10%	5-10%

[1]Wang et al. (2015), [2] Liu et al. (2019), [3] Imdahl et al. (2020), [4]Blattman et al. (2020), [5]Kuchina et al. (2021)

1.7 Summary and objectives

SCG and SCT have tremendous potential to bring more clarity to the nature of MDM and their metabolic potentials, which will enable us to provide information on individual organisms and the structure and dynamics of natural microbial populations in various environments. SC 'omics holds great promise in microbial microevolution studies, industrial bioprospecting, and selection of suitable heterologous expression systems, with potential for novel and environmentally responsible energy solutions, bioremediation of toxins, and natural products (Kaster & Sobol, 2020). Undoubtedly, the future for SC 'omics is exceptionally bright, but significant technical and conceptual challenges still have to be resolved. Solving these problems will widen this fields reach to diverse microorganisms and will help create a more phylogenetically balanced representation of genomes in databases. In turn, this will ultimately help to improve models for computational gene annotation and taxonomic assignment (de Jager & Siezen, 2011; Y. Wang & Navin, 2015; Woyke et al., 2009) and help with the cultivation of currently unculturable microorganisms by revealing their nutritional needs and metabolic capabilities (Pratscher et al., 2018). Therefore, the goals of this doctoral work were to improve upon existing methodology in both microbial SCG and SCT to further advance the study of microbial single-cells.

Chapter 1: Improving microbial single-cell genomics

In **Chapter 1**, the overall objective was to improve SCG, specifically the cell labeling, isolation, lysis, and amplification steps. First, a FISH approach for targeted-cell labelling compatible with FACS and downstream amplification methods was improved. This enabled the analysis of rare microbial taxa representing less than 1% within a complex environmental community, which would have otherwise gone undetected with by metagenomics (Dam et al., 2020). In the second and third task, an optimized cell sorting and subsequent cell lysis step were established to better avoid premature cell lysis and lessen DNA damaged caused by lysis buffers to improve genome recovery. Lastly, MDA reaction volumes were sequentially reduced to the sub-microliter range in 384-well plates with the help of a non-contact liquid dispenser and amplification bias was compared. From this, it could be determined that genome coverages >90% and more uniform amplification could be achieved in 1.25 µL reaction volumes without the need for specialized microfluidic equipment, which further reduced complexity and costs associated with SCG.

Chapter 2: Improving microbial single-cell transcriptomics

The goal of **Chapter 2** was to establish a SCT pipeline for prokaryotes that could reliable detect transcripts from single cells, while also providing insight into their heterogeneity in an isogenic population. First, the challenges and limitations of working with low levels of RNA were assessed by benchmarking a standard reverse transcription approach on ≤ 1000 *E. coli* cells. Next, an attempt to establish a novel, amplification-free SCT method using Oxford Nanopore sequencing kits was made, but the method was found to not be sensitive enough for low concentrations of RNA. Finally, microbial SCT was accomplished by modifying an MDA-based single-cell WTA method. With this improved approach, differential treatment between heat-shocked and non-treated cells could be determined from the scRNA-seq data. Furthermore, rare functions could be identified that were overlooked with respective bulk RNA-seq.

2 Materials and Methods

2.1 Chapter 1: Improving microbial single-cell genomics

2.1.1 Targeted-cell sorting with fluorescence *in situ* hybridization

The first objective of Chapter 1 was to improve upon the cell labeling approach so that specific taxa could be enriched during downstream cell isolation. Prior to this approach, standard labeling involved non-specific DNA stains and was therefore less efficient. Here, fluorescence *in situ* hybridization (FISH) was modified to be compatible with fluorescence-activated cell sorting (FACS) and multiple displacement amplification (MDA) by enabling FISH to be performed in solution and without paraformaldehyde fixative (Dam et al., 2020). An environmental sample from a winery wastewater treatment plant (WWTP) was chosen for benchmarking. In this sample, the bacterial phylum Chloroflexi, representing < 1% of the total community, was selected as the target to prove that the in-solution, fixation-free FISH was sensitive enough to target low abundant cells. Additionally, Chloroflexi was chosen since this phylum is a deep-branching lineage which exhibits a wide range of metabolic activities (Hug et al., 2013; Islam et al., 2019). They are also of high interest for the biotech industry as many members are estimated to produce novel antimicrobials (Dam et al., 2020; Hemmerling & Piel, 2022; Kogawa et al., 2022). Metagenomics and 16S rRNA data show that Chloroflexi are ubiquitous throughout the environment, however, they have also mostly evaded cultivation attempts and are therefore under-characterized. Thus, by enriching low abundant Chloroflexi from an environmental sample for single-cell genomics it was anticipated that rare and/or novel taxa could be identified that would have otherwise been overlooked with metagenomics.

2.1.1.1 Sample collection and preparation

Wastewater samples from the aerated lagoon (LEA) of the WWTP of the Establecimiento Juanicó winery (located in the village Juanicó in Canelones, Uruguay, latitude -34.6, longitude -56.25) were collected 20 cm below the water level. The samples were vortexed at maximum speed for 3 min to release cells attached to the sediments. After 1 hr the sample was centrifuged at 2,500 rpm for 30 sec to remove large particles (Rinke et al., 2014). Supernatant was filtered through a 30 µm Celltrics® Filter to further remove large particles (Sysmex, Germany).

2.1.1.2 In solution, fixation-free fluorescence *in situ* hybridization

Cells in the sample were hybridized with equal amounts of two probes labeled with Cyanine3 fluorochrome that target the phylum Chloroflexi: GNSB941 (5'-AAACCACACGCTCCGCT-3') (Gich et al., 2001) and CFX1223 (5'-CCATTGTAGCGTGTGTGTMG-3') (Björnsson et al., 2002). The hybridization protocol used in this study was modified from the protocol by (Yilmaz et al., 2010) and (Pernthaler et al., 2001) as follows: Cells were pelleted, washed twice with 1X phosphate buffered saline (PBS) to remove possible fluorescent molecules, and hybridized with the two probes, each at a final concentration of 15 ng μL^{-1} in 100 µL hybridization buffer containing 35% formamide at 46°C for 3 h in the dark. Labeled cells were washed twice with pre-warmed wash buffer at 48°C for 20 min each. Cells were then washed for the last time with ice cold 1X PBS buffer before being re-suspended in 500 µL 1X PBS buffer. The negative “no-probe” control was treated the same way as labeled samples except that no probes were added during hybridization step. To test if the fluorescence signal of hybridized cells could be improved, cells were treated with increasing concentrations of ethanol (50%, 80%, and 98%) with 3 min incubation times (Haroon et al., 2013). Hybridized cells were visualized with an Axiophot fluorescence microscope (Carl Zeiss Microimaging GmbH). Labeled cells were stored in 5% glycerol at -80°C for sorting the next day with no loss of signal. In order to verify the specificity and sensitivity of labeling Chloroflexi, a mixed culture containing 1% *Sphaerobacter thermophilus* (DSM20745) and 99% *Escherichia coli* K12 (DSM498) was used. The hybridization procedure was carried out as described with the WWTP samples.

2.1.1.3 Targeted cell sorting of labeled cells

Cell sorting of labeled cells was performed using a BD FACSARIA III cell sorting system (BD, Germany). A 488 and 561 nm laser were used as excitation source for light scattering and fluorescence, respectively. Hybridized cells were diluted 5X in 1X PBS, filtered through a 10 µm Celltrics® filter (Sysmex, Germany), and briefly sonicated in an Ultrasonic cleaner (VWR, Germany) to break up potentially aggregated cells. Labeled cells were enriched by sorting into a 5 mL Falcon® polypropylene tube (Corning, USA) using purity sort mode. Cells from the enrichment sort were sorted into Hard-Shell® 384-well plates (Bio-Rad Laboratories, Germany) using the single cell mode of the FACS at a lower speed (50–100 cells sec^{-1}). Cells were then sorted based on signal intensity of forward scattering and emitted fluorescence, compared to those of the no-probe control.

2.1.1.4 Multiple displacement amplification

Cells were lysed and their genomic DNA was released during alkaline lysis supplied by the REPLI-g® Single Cell Kit (Qiagen, Germany) at 65°C for 10 min. Genomic DNA was amplified with phi29 DNA polymerase at 30°C for 6 h using the REPLI-g® Single Cell Kit (Qiagen, Germany) on a CFX384 Touch™ Real-Time Detection System (Bio-Rad Laboratories, Germany). Amplification was monitored in real time by detection of SYTO13® (Life Technologies, USA) fluorescence every 5 min. MDA reactions were then terminated at 65°C for 10 min. The cycle quantification (Cq) values and endpoint relative fluorescence units (RFU) were used to determine the positive amplifications.

2.1.1.5 16S rRNA gene amplification and screening

MDA products were diluted 1:20 and used as templates to amplify 16S rRNA genes with universal bacterial primer pairs: 926wF: 5'-AAACTYAAAKGAATTGRCGG-3' and 1392R: 5'-ACGGGCGGTGTGTRC-3' (Rinke et al., 2014). PCR products were cleaned up with DNA Clean and Concentrator-5 (Zymo Research, Germany) and subjected to Sanger sequencing. 16S rRNA gene sequences were blasted against the Silva SSU database (version 132, released in December 2017) and the identities of the corresponding single cells were determined using the web-based tool SINA Search and Classify on www.arb-silva.de (Pruesse et al., 2012).

2.1.1.6 DNA extraction for metagenome sequencing

DNA from the WWTP samples was extracted using a hexadecyltrimethylammonium bromide (CTAB)-based method (R. I. Griffiths et al., 2000) with some modifications as follows: 1.5 mL of the samples were centrifuged at maximum speed for 5 min to collect biomass. Pellets were then transferred into a Lysing matrix E beads (MP Biomedicals, France). 500 µL 6% CTAB extraction buffer and 500 µL phenol:chloroform:isoamyl (PCI) alcohol (25:24:1) were added into the extraction tube. Cells were lysed by vortexing at maximum speed on a Vortex Genie2 (Scientific Industries, USA) for 3 min. The supernatant was extracted twice with PCI (25:24:1) and twice with chloroform:isoamyl alcohol (24:1). The aqueous phase was transferred into a clean 1.5 mL tube. DNA was precipitated with 2.5X volume of 100% ethanol and 0.1X volume of 3 M sodium acetate (pH 5.2) and re-suspended in 50 µL PCR grade water. Extracted DNA was cleaned up with the DNA Clean and Concentrator-5 kit (Zymo Research, Germany) as per the manufacturer's instruction. Preliminary survey of microbial communities in WWTP samples were performed using pyrosequencing.

2.1.1.7 Library preparation for metagenome and single cell genome sequencing

Genomic DNA extracted from the WWTP samples and MDA products was quantified using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific, USA). Libraries were prepared using the NEBNext® Ultra™ DNA Library Prep Kit and NEBNext® Ultra™ II FS DNA Library Prep Kit (New England BioLabs, Germany), respectively, following the manufacturer's instruction. 500 ng of DNA was used as starting material. The quality of the DNA libraries was verified using the Agilent High Sensitivity DNA Kit on the Agilent 2100 Bioanalyzer instrument (Agilent Technologies, Germany). The libraries were then pooled and sequenced on Illumina systems using the paired-end approach and the highest available read length for each platform (150 bp for NovoSeq and NextSeq, 300 bp for MiSeq).

2.1.1.8 Read processing and assembly

Quality trimming and adapter clipping was done using a three-step process, consisting of Trimmomatic v.0.36 (Bolger et al., 2014), bbduk v.35.69 (Bushnell, 2014) and cutadapt v.1.14 (Martin, 2011) using the following argument settings, respectively:

Trimmomatic: "ILLUMINACLIP: Trueseq3_PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:80"
Bbduk: "-ktrim=r -mink=11 -minlength=45 -entropy=0.25"
Cutadapt: "-a AGATCGG$ -a CCGATCT$ -A AGATCGG$ -A CCGATCT$"

Overlapping read pairs were identified and merged using FLASH v.1.2.11 (Magoč & Salzberg, 2011) with a minimum overlap of 16 bp, a maximum overlap of 100 bp and a maximum mismatch fraction of 0.1. Residual contaminants of the Illumina PhiX control spike-in were removed using fastq_screen v.0.4.4 (Wingett & Andrews, 2018). All datasets were assembled with SPAdes v.3.10.1 (Nurk et al., 2013), iterating through kmers 21- 121 with a step-size of 10 and using the "careful" argument. The "--sc" flag was used for all single cell datasets, while the "--meta" flag was used for metagenome datasets. Winery metagenome samples obtained from different years were assembled individually and then subsequently merged using minimus2 (Sommer et al., 2007).

2.1.1.9 Genome assessment and co-assembly

Genome completeness and purity was assessed using checkM (Parks et al., 2015). For taxonomic assignment, additional purity assessments as well as for decontamination purposes, a hierarchical least common ancestor (LCA) contig classification approach was performed as described by (Pratscher et al., 2018), using preliminary assignments based on 16S rRNA, 23S rRNA, universal single copy marker genes, as well as total protein sequences. Contigs with confident hierarchical taxon assignments that conflicted with the predominant taxon classification of the respective genome were removed as potential contaminations. The average nucleotide identity (ANI) approach implemented in pyani v.0.2.7 (Pritchard et al., 2016) was employed to identify groups of SAGs belonging to the same species using an identity cutoff of

≥99% and a coverage cutoff of 10%. SAGs of the same species were merged and reassembled into co-assembled genomes (CAGs). SAGs with a genome coverage of less than 5% were omitted from analysis.

2.1.1.10 Coverage assessment and binning

Metagenome coverage of all SAG, CAG, and merged metagenome contigs were obtained by mapping reads back to the assemblies using BamM v.1.7.3 (Woodcroft et al., 2019). MAGs were obtained via metagenome binning by combining the results obtained from Maxbin v.2.2.6 (Wu et al., 2016), CONCOCT v.1.0.0 (Alneberg et al., 2013), as well as MetaBat v.2.12.1 (Kang et al., 2015) using DAS Tool v1.1.1 (Sieber et al., 2018). SAGs which, based on CheckM (Parks et al., 2015) evaluations and marker-gene phylogenies, potentially consisted of multiple co-sorted cells, were separated into the respective potential component genomes by binning using Maxbin v.2.2.6 together with metagenome coverage information. After each binning and re-assembly step, the completeness and purity of all bins and SAGs were re-assessed using checkM as well as the hierarchical contig classification procedure described in (Pratscher et al., 2018).

2.1.1.11 Phylogenetic analysis of Chloroflexi genomes

Primary taxonomic assignments were inferred from the hierarchical contig classification results obtained during genome assessment. For comparison purposes, additional assignments were inferred using GTDB-TK (Chaumeil et al., 2019). 16S rRNA phylogenies were reconstructed using the Arb software package (Westram et al., 2011), which aligned 16S rRNA gene sequences amplified from Chloroflexi SAGs and CAGs, as well as selected reference Chloroflexi isolates. *Streptomyces griseus* was used as the outgroup. A phylogenetic tree was inferred using the neighbor joining algorithm with 1000 bootstrap permutations.

Proteinortho5 (v.5.16b) (Lechner et al., 2011) was used to detect groups of orthologous genes shared between reference genomes and CAGs, SAGs, and MAGs in our study with the following parameters: -identity=25 -e=1e-10 -cov=60 -selfblast -singles. A gene-content based genome clustering based on the presence or absence of genes from the bidirectional blast results of Proteinortho was implemented with a custom python script (https://npm.pkg.github.com/jvollme/PO_2_GENECONTENT) using the neighbor joining algorithm with 1000 bootstrap permutations. *Streptomyces griseus* was also used as an outgroup.

2.1.1.12 Genome analysis

Preliminary gene calling and annotations were inferred using different platforms including Prokka pipeline v1.12-beta (Seemann, 2014), Rapid Annotations using Subsystem Technology (RAST) (Aziz et al., 2008), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2021). AntiSMASH (v4.1.0) (Blin et al., 2017) was used to identify putative secondary metabolite gene clusters.

2.1.1.13 Pyrosequencing

DNA was extracted using the ZR Soil Microbe DNA MiniPrepTM (Zymo Research, US) as described per the manufacturer instructions. DNA was dehydrated with 95% ethanol and submitted to the Institute for Agrobiotechnology Rosario (INDEAR, Rosario, Argentina) for 454-pyrosequencing and bioinformatic analysis (Roche Genome Sequencer FLX Titanium system). For sample LEA2013 the 16S rRNA genes were amplified with primers for the V4 region: 563f (5'-AYTGGGYDTAAAGNG-3') and 802r (CAGGAAACAGCTATGACC) using a 10 bp barcode. For samples LEA2014 and LEA2015 the 16S rRNA genes were amplified with primers for the V3- V4 regions: 357F (5’-CACGACGTTGTAAAACGACCCTACGGGAGGCAGCAG-3’)/926R (5’-CAGGAAACAGCTATGACCCCGTCAATTCMTTTRAGT-3’) using a 10 bp barcode. Sequences were analyzed using the Quantitative Insights Into Microbial Ecology (QIIME) software (Caporaso et al., 2010).

Reads with length less than 200 bases, quality coefficient greater than 25, homopolymer size higher than six and ambiguous bases were removed. Operational Taxonomic Units (OTU) were defined using UClust algorithm based on 97% identity, OTUs that contained less than one sequences (singletons) were removed from the analysis. Reads were classified using the Classifier tool, from the Ribosomal Database Project (http://edp.cme.msu.edu/classifier/classifier.jps) with a cutoff of 50%.

2.1.2 Improving cell isolation with a single-cell printer

In order for cells to be successfully lysed and amplified after the sorting process, they must first survive the isolation process. There are several types of isolation methods (discussed in Section 1.5.3), which all have their disadvantages and advantages. Yet, FACS has remained the most commonly used technology for sorting microbial single cells. This is due mainly to the fact that FACS is very high-throughput and is sensitive towards fluorescently labeled cells (Blainey, 2013; Rinke et al., 2014). However, one of the main known issues with FACS is its high sorting pressure which prematurely lyse damaged and/or sensitive cells (Blainey, 2013; Mollet et al., 2008; Wiegand et al., 2020). Therefore, the second objective of Chapter 1 was to compare the viability of Gram-negative and Gram-positive cells pre-treated with FISH after FACS and single-cell printing (SCP). In this study, Gram-negative *Flavobacterium denitrificans* (DSM 15936) and Gram-positive *Bacillus marisflavi* TF11 (DSM 16204) were chosen as the text subjects instead of *E. coli* because of the differences in their genomic GC content and in cell wall structures. *F. denitrificans* is easier to lyse and represents low GC at ~37%, whereas *B. marisflavi* is harder to lyse and has an average GC of 49%. Becaus*e* the SCP does not apply high-pressure during sorting, it was hypothesized that fewer cells damaged by FISH would viable after FACS.

2.1.2.1 Bacterial growth and FISH

F. denitrificans and *B. marisflavi* were grown in 3 mL of Luria Bertani (LB) and Bacto Marine Broth, respectively, at 28°C and 180 rpm in an Infors HT Multitron incubating shaker (Infors AG, Switzerland) until exponential phase e.g. optical density (OD) 600 of 0.8-1.2. The cells were centrifuged at room temperature at 10,000 x g for 3 min. The supernatant was discarded and the cell pellet washed by resuspension in 500 µL of 1X PBS. The washing step was repeated once. FISH was applied like above, except the probes LGC345a, b, and c (Meier et al., 1999) were used to label *B. marisflavi* and the probe CF319A was used to label *F. denitrificans* (Manz et al., 1996). Additionally, cells treated with FISH, but not provided probe were also prepared as a negative control.

2.1.2.2 Labeled cell-sorting

A BD FACSMelody (Becton-Dickson, USA), fitted with a 100 µM nozzle and equipped with a 488nm laser for excitation was used to sort cells. Cells were first diluted to approximately 10^6 cells mL^{-1} with sterile 1X PBS to ensure an event rate of <1000 events/s. Gates were defined on side-scatter (cell complexity) and forward-scatter (cell-size), not fluorescence of the FISH probe as the cell printer (described in detail below), was not able to sort based on fluorescence. Triplicates of 10 and 100 cells were sorted into 96 well plates (Bio-Rad Laboratories, USA) with 200 µL of the respective medium inside.

Additionally, the same diluted and filtered labeled cells were also sorted into the same plates using the cell printer B.SIGHT™ (Cytena, Germany). Cells were isolated based on a size of 2 µM to 4 µM and a roundness of 0.6 to 1.0 with bright-field. The plates were incubated for 5 days at 28°C and 180 rpm in an Infors HT Multitron incubating shaker (Infors AG, Switzerland). The viability (growth) after 5 days was measured with the Spectramax M2 (Molecular Devices, USA).

2.1.3 Improving cell lysis and multiple displacement amplification

After successful cell isolation in 384 well plates the cells are lysed (**Figure 1.2**). Lysis must be efficient so that DNA is released but not damaged and fragmented. Highly fragmented DNA cannot be primed and synthesized by polymerases used in WGA that require large templates (Blainey, 2013). Since microbial cells have very diverse morphologies, adapting a “one-size-fits-all” lysis solution is difficult. When possible, lysis should be adapted to the sample for better WGA success. Therefore, a modified lysis was adapted for the model organism *E. coli* used herein.

The final objective for Chapter 1 was to improve upon the WGA method MDA. MDA has remained one of the most commonly used microbial WGA methods, but the bias MDA has against high GC content and the overamplification of some genome regions has often made it difficult to obtain high quality single-cell genomes. Furthermore, the high risk of contamination and large costs have also limited the accessibility of MDA. Therefore, the goal of this work to find an approach that reduced costs and contamination, while improving genome coverage and coverage uniformity. It was hypothesized that lowering MDA reaction volumes using standard and commercially available SCG equipment would provide an easy solution to this problem.

2.1.3.1 Bacterial cell isolation

Escherichia coli K12 MG1655 (DSMZ 18039) was cultured in 1 mL of Luria Bertani (LB) broth at 30°C and 750 rpm with the Thermomixer Comfort (Eppendorf, Germany) to exponential growth phase (OD600 of ~ 2.2-2.6). From this point forward, cells were processed in a UV decontaminated ISO 4 cleanroom. Equipment and gloves were decontaminated with DNA AWAY (Thermo Fisher Scientific, USA). Consumables were UV treated for 1 hr in a crosslinker and 1X PBS was UV treated for 6 hours in a 254nm shortwave ultraviolet crosslinker at 0.12 Joules/cm^2 (Analytik Jena US).

A BD FACSMelody (Becton-Dickson, USA), fitted with a 100 µM nozzle and equipped with a 488nm laser for excitation was used to sort single cells. Cells were first diluted to approximately 10^6 cells mL^{-1} with sterile 1X PBS to ensure an event rate of <1000 events/s. Gates were defined on side-scatter (cell complexity) and forward-scatter (cell-size). Cells were sorted in single-cell mode into 384-well plates (Bio-Rad, USA) containing no sorting buffer (i.e. dry sorting). Plates were sealed with Microseal B (Bio-Rad, USA) and stored at -80°C.

2.1.3.2 Cell lysis

Plates containing sorted cells were thawed and centrifuged at 4°C for 5 min at 3000 rpm (Eppendorf, Germany). WGA-X ™ cell lysis buffer consisting of 0.4 M KOH, 10 mM EDTA and 100 mM and 1M Tris-HCl neutralization buffer (Stepanauskas et al., 2017), were treated with UV for 10 min on an ice-water bath in a 254 nm shortwave ultraviolet crosslinker at 0.12 Joules/cm^2 (Analytik Jena US) (Woyke et al., 2011). A volume of 0.088 µL of lysis solution containing either 0.5 M KOH, 10 mM EDTA, and 100 mM dithiothreitol (DTT) (100% lysis buffer) or 0.2 M KOH, 5 mM EDTA, and 50 mM DTT (50% lysis buffer), were dispensed with an I.DOT mini (Dispendix, Germany) non-contact liquid dispenser. Additionally, there was a set of control cells that did not receive lysis solution. Lysis was incubated at 21°C for 10 min and neutralized by the addition of an equal volume of 1M Tris-HCL, pH 4. A comparison between the different lysis conditions found better genome recovery with 50% lysis buffer, which was used in subsequent, optimized reactions. The amount of lysis and neutralization buffer per a MDA reaction can be found in **Table 2.1**.

Table 2.1. MDA reagents by reaction volume

MDA Reaction Volume	Lysis Buffer	Neutralization Buffer	H_2O	REPLI-g sc Reaction Buffer	REPLI-g sc Polymerase	Syto-13
0.8 µL	0.056	0.056	0.175	0.464	0.032	0.016
1.0 µL	0.070	0.070	0.219	0.580	0.040	0.020
1.25 µL	0.088	0.088	0.274	0.725	0.050	0.025
5.0 µL	0.350	0.350	1.095	2.900	0.200	0.100
10 µL	0.700	0.700	2.190	5.800	0.400	0.200

2.1.3.3 Multiple displacement amplification

Multiple displacement amplification (MDA) was performed with the REPLI-g Single Cell Kit (QIAGEN, Germany). REPLI-g sc Reaction Buffer and Polymerase were combined in 0.2 mL DNase, RNase-free PCR tubes (Biozym Scientific GmbH, Germany) and UV treated for 30 min on an ice-water bath in a 254 nm shortwave ultraviolet crosslinker at 0.12 Joules/cm^2 (Analytik Jena US) (Woyke et al., 2011). SYTO™-13 (Invitrogen, USA) was added to the master mix at a final concentration of 1 µM. SYTO™-13 is used to monitor the progression of MDA since it binds to newly formed double-stranded DNA as it is amplified. The REPLI-g master mix was then dispensed onto the lysed cells with an I.DOT mini (Dispendix, Germany) non-contact liquid dispenser so that the final MDA volumes were 0.5 µL, 0.8 µL, 1.0 µL, 1.25 µL, 5 µL, and 10 µL (**Table 2.1**). The MDA's were incubated for 6 hours at 30°C in a CFX-384 thermocycler (Bio-Rad, USA), then 65°C for 10 min to stop the amplification and held at 4°C. Amplified DNA was kept at -20°C until used for library preparation.

2.1.3.4 Library preparation and sequencing

The following steps were performed under a UV decontaminated Laminar Flow PCR workbench (STARLAB International GmbH, Germany), sterilized with DNA AWAY (Thermo Fisher

Scientific, USA). Prior to library preparation, the amplified DNA was cleaned with DNA Clean & Concentrator – 5 (Zymo Research, USA). DNA input for library preparation was normalized to 5.98 ng/µL. Libraries were prepared using the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina (New England Biolabs (NEB), USA), following the <100 ng input protocol. Fragmentation was set to 14 min and 7 PCR cycles were used. NEBNext® Multiplex Oligos for Illumina® was used for barcoding. Library concentration and size was quantified with Qubit™ DNA HS assay (Life Technologies, USA) and Bioanalyzer High Sensitivity DNA kit (Agilent, USA). The libraries were sequenced using an Illumina NextSeq 550 with the High Output Kit v2.5 -300 Cycles (2 x 150 bp paired-end) (Illumina, USA).

2.1.3.5 Data processing and analysis

The sequence reads were quality checked using FastQC v0.11.9 (www.bioinformatics.babraham.ac.uk/projects/fastqc) and quality-trimmed using Trim Galore (Krueger et al., 2021). Following trimming, reads were normalized to 3,108,153 read pairs with BBTools reformat.sh (Bushnell, 2014). Normalized reads were assessed for contamination using FASTQ-Screen v0.15.2 (Wingett & Andrews, 2018), *E. coli* multi-mapping reads were kept. PCR duplicates were counted and removed with *dedupe.sh* from BBTools (Bushnell, 2014). Then reads were mapped to *E. coli* MG1655 (ASM584v2) with *bbmap.sh*. Max indel length was set to 80, as recommended for MDA, then coverage was calculated for 1kb bins (Bushnell, 2014).

The three replicates with the lowest read coverage standard deviation in 10 kb bins were chosen for each sample type for assembly. Prior to *de novo* assembly, the read coverage was normalized with *bbnorm.sh* with target=100 and min=5 (Bushnell, 2014). SPAdes v.3.15.5 was used as recommended for single-cells by using the flag –sc for single-cell mode, kmer lengths of 21 to 101 in 10 step increments, and setting the flag –careful to reduce the number of mismatches (Prjibelski et al., 2020). QUAST v.5.2.0 was used to assess assembly quality (Mikheenko et al., 2018) and MDMcleaner (Vollmers et al., 2022) was used to check for contamination and completeness.

Statistical differences between sample quality, mapping and assembly statistics were calculated using Anova: Single Factor with an alpha value of 0.05 in Microsoft Excel®. Gini indexes were calculated with the ineq package (Zeileis et al., 2014) in R v.3.6.3 (R Core Team, 2015). Read depth and Lorenz curve plots were created using ggplot2 (Villanueva & Chen, 2019).

2.1.4 MDA test on a droplet microarray

MDA reaction volume reduction was also tested on the droplet microarray (DMA), which is a platform consisting of a glass side with super-hydrophobic and hydrophilic patterning that creates spots for nanoliter reactions to take place (Feng et al., 2018; Jogia et al., 2016). Because the droplets only have contact with one surface, in theory, nonspecific surface adsorption of nucleic acids should be minimized (Belotserkovskii et al., 1996; Gaillard & Strauss, 1998). Thus, the DMA chip was used to further reduce reaction volumes and compare the results to MDA performed in the 384 multi-well plates.

2.1.4.1 Reducing evaporation issues with glycerol

Glycerol's effect on MDA was fir benchmarked with single *E. coli* cells that were sorted into 384-well plates with the BD FACSMelody (Becton-Dickson, USA) and treated with 50% lysis buffer and 1M Tris-HCL neutralization buffer, as mentioned above. Master mix, supplemented with final concentrations of 5% and 10% glycerol, were dispensed onto the cells in either 5 µL of 1.25 µL MDA reaction volumes. Comparisons between the amplification of 5 or 10% glycerol MDA was monitored with the CFX384 Touch Real-Time PCR thermocycler (Bio-Rad, USA). Because 5% MDA master mix gave better results, it was used in further experiments.

2.1.4.2 MDA on the DMA

DMAs were sterilized with 70% isopropanol and UV treated for 5 min in a 254nm shortwave ultraviolet crosslinker at 0.12 Joules/cm^2 (Analytik Jena US). A humidity chamber and humidified petri dish meant to reduce evaporation off the DMA during the long-term amplification incubations (Chakraborty et al., 2022), were prepared 30 min prior to MDA. For DNA testing, heat-denatured (95°C for 2 min) DNA was dispensed with the I.DOT (Dispendix, Germany). Because the BD FACSMelody cannot sort onto the DMA, the B.SIGHT™ (Cytena, Germany) was used for single-cell sorting onto the DMA. Cells were prepared and isolated under the same settings as mentioned above. During the room temperature lysis incubation, the DMA was kept in the humidified petri dish. MDA master mix was decontaminated with UV for 30 min as described above and dispensed with the I.DOT so that the final volume was 0.5 µL. Initially, MDA incubations for 6 hours were performed in a C1000 Touch thermal cycler fitted with a 96-well block (Bio-Rad, USA) which held the humidity chamber that the DMA was placed in. However, it was later determined that performing the MDA incubation in an incubator instead at 32°C with the DMA inside the humidified petri dish reduced evaporation. After amplification, 1 µL of single-cell grade H_20 was dispensed onto each reaction. The reactions were quantified with Qubit™ DNA HS assay (Life Technologies, USA).

2.2 Chapter 2: Improving microbial single-cell transcriptomics

2.2.1 Improving reverse transcription for low cell inputs

The first task of Chapter 2 was to gauge what the RNA input limits were for two standard reverse transcription methods. This was done beginning with total *E. coli* RNA diluted to concentrations equivalent to 1 million cells down to 1000 cells. Then, the superior method was validated with RNA extracted from 1000 and 100 sorted *E. coli* cells.

2.2.1.1 Bacterial growth conditions and cell isolation

Escherichia coli K12 MG1655 (DSMZ 18039) was cultured in 1 mL of LB broth at 30°C and 750 rpm with the Thermomixer Comfort (Eppendorf, Germany) to exponential growth phase (OD600 of ~ 2.2-2.6). For tests with sorted cells, cells were processed in a UV decontaminated ISO 4 cleanroom. Equipment and gloves were decontaminated with DNA AWAY (Thermo Fisher Scientific, USA). Consumables were UV treated for 1 hr in a crosslinker and 1X PBS was UV treated for 6 hours in a 254nm shortwave ultraviolet crosslinker at 0.12 Joules/cm^2 (Analytik Jena US). Prior to sorting, cells were diluted to ~ 10^6 cells mL^{-1} with sterile 1X PBS. A BD FACSMelody (Becton-Dickson, USA), fitted with a 100 µM nozzle, was used to sort cells into 2 mL tubes containing 250 µL of RNAPure™ peqGOLD (VWR International GmbH, Germany), then immediately vortexed and kept on ice until extraction.

2.2.1.2 RNA extraction

For tests that used diluted bulk RNA, cells were first washed and resuspended with 100 µL of 1X PBS. The 100 µL sample was then transferred to a Lysing matrix E bead tube (MP Biomedicals, France) and supplied with 300 µL of RNAPure™ peqGOLD (VWR International GmbH, Germany). Both bulk RNA and sorted cell RNA were extracted with the Direct-zol RNA Miniprep Kit (Zymo Research, USA), following the standard protocol. DNase treatment was applied after extraction using the TURBO DNA-free™ kit (Invitrogen, USA) and the RNA stored in 1 µL of RNasin® Ribonuclease Inhibitor (RI) (Promega Corporation, USA). RNA was quantified with the Qubit™ RNA HS Assay Kit (Invitrogen, USA). RNA was diluted to 100 ng µL^{-1}, 10 ng µL^{-1}, 1 ng µL^{-1}, and 0.1 ng µL^{-1} from bulk extracted RNA prior to storage.

2.2.1.3 Reverse transcription and PCR

First, 1 µL of RNA and 1 µL of 200 pg Exo-Resistant random primer (Thermo Scientific, USA) were combined in 0.2 µL PCR tubes (Biozym, Germany) and denatured by heating the samples to 70°C for 5 min. The samples were then immediately placed on ice. Then, 18 µL of reverse transcription master mix was added to the samples. For tests with MMLV (Promega, USA) the final reagent concentration in the master mix was as follows: 1.25X MMLV Reaction Buffer (Promega, USA), 0.625 mM dNTP (NEB, USA), 40U RNasin RI (Promega, USA), and 87.5 U MMLV reverse transcriptase (Promega, USA). For tests with SuperScript IV (ThermoFisher, USA) the final concentrations are as follows: 1.25X SuperScript IV Reaction Buffer (ThermoFisher, USA), 0.625 mM dNTP (NEB, USA), 40U RNasin RI (Promega, USA), and 35U SuperScript IV reverse transcriptase (ThermoFisher, USA). RT was incubated at 23°C for 10 min and either 42°C (MMLV) or 50°C (SuperScript IV) for 1.5 hrs in a CFX96 Touch PCR thermocycler (Bio-Rad, USA. RT was terminated at 80°C for 10 min.

Next, cDNA was amplified with PCR. The final concentrations for PCR reagents included: 2X Standard Taq Reaction Buffer (NEB, USA), 0.2 µM of both 926wF: 5'-AAACTYAAAKGAATTGRCGG-3' and 1392R: 5'- ACGGGCGGTGTGTRC-3' primers, 0.2 mM dNTP (NEB, USA), and 0.125U Taq Polymerase (NEB, USA). 2 µL were of the RT reaction was used as template. The thermocycling conditions were: 94°C for 5 min, 30 cycles of 94°C 30s, 52°C 40s, and 68°C for 45 s, final extension was 68°C for 5 min. Successful RT was analyzed with gel electrophoresis using 1% agarose (Bio-Rad, USA) at 80V for 45 min.

2.2.2 Oxford Nanopore direct-cDNA sequencing

The next step in Chapter 2 was to develop a novel microbial SCT method. First, amplification-free RNA-seq with Oxford Nanopore's direct-cDNA kit was tested to determine if the method was sensitive enough for < 100 ng of total RNA. This approach was selected because removing amplification improves the reliability for transcript quantification (Parekh et al., 2016). However, because this method requires RNA with polyA tails, a polyadenylation step was implemented before library preparation and sequencing.

2.2.2.1 Bacterial growth and RNA extraction

Escherichia coli K12 MG1655 (DSMZ 18039) was cultured in 1 mL of LB broth at 30°C and 750 rpm with the Thermomixer Comfort (Eppendorf, Germany) to exponential growth phase (OD600 of ~ 2.2-2.6). The cells were then washed and resuspended with 100 µL of 1X PBS. The 100 µL sample was then transferred to a Lysing matrix E bead tube (MP Biomedicals, France) and supplied with 300 µL of RNAPure™ peqGOLD (VWR International GmbH, Germany). RNA was extracted with the Direct-zol RNA Miniprep Kit (Zymo Research, USA), following the standard protocol. DNase treatment was applied after extraction using the TURBO DNA-free™ kit (Invitrogen, USA) and the RNA stored in 1 µL of RNasin® RI (Promega Corporation, USA). RNA was quantified with the Qubit™ RNA HS Assay Kit (Invitrogen, USA).

2.2.2.2 Polyadenylation

The polyadenylation protocol was modified from (Grünberger et al., 2022; Wongsurawat et al., 2019) which uses *E. coli* Poly(A) Polymerase (NEB, USA). The total reaction volume was reduced from 50 µL to 26 µL. RNA was first incubated at 70°C for 2 min, then kept on ice. Next, 20U of poly(A) polymerase (NEB), 2 µL of 10X reaction buffer (NEB), 4.5 µL of 10 mM ATP (NEB), and 0.5 µL of RNasin RI (Promega, USA) was added to the denatured RNA. The polyadenylation reaction was incubated at 37°C for 10 min. AmpPureXP beads (2.5X) were used to collect the RNA from the reaction. RNA was quantified with the Qubit™ RNA HS Assay Kit (Invitrogen, USA) and with the Agilent High Sensitivity RNA Kit on the Agilent 2100 Bioanalyzer instrument (Agilent Technologies, Germany). RNA was stored overnight at -80°C in 1 µL of RNasin RI.

2.2.2.3 Library preparation and RNA sequencing

RNA libraries from 40 and 4 ng of total RNA were prepared with the ONT Direct-cDNA kit (ONT, UK). Modifications to the standard user protocol (SQK-DCS109 with EXP-NBD104) are as follows: from the RNA degradation and second strand synthesis step, all volumes were reduced to ½ the original volume, end prep was increased to 10 min, and bead washing steps were with 80% ethanol instead of 70%. Final libraries were quantified with Qubit™ DNA HS assay (Life Technologies, USA) and Bioanalyzer High Sensitivity DNA kit (Agilent, USA). The libraries were sequenced with a MinION using a Flongle flow cell.

2.2.2.4 Data processing and analysis

Fast5 files were base called and trimmed using Guppy v4.5.2 (ONT, UK). FASTA files were mapped to *E. coli* MG1655 strain (ASM584v2, 28-11-2021) genome with minimap2 (Li, 2018). Transcripts were counted with featureCounts (Liao et al., 2014). Finally, coverage information was calculated with package NOIseq (Tarazona et al., 2012) in R v4.0.0 (R Core Team, 2015).

2.2.3 Modification of the REPLI-g single-cell WTA kit for prokaryotes

Since the ONT direct-cDNA results were not successful, the final step was to establish a functioning microbial SCT method. Due to the familiarity with MDA, the REPLI-g WTA Single Cell kit (QIAGEN, Germany) was chosen. However, the kit is originally designed for eukaryotes and claims that less than 10 pg, which is ~100X more than a single *E. coli* cell. Therefore, it was presumed that modifications to the protocol could successfully produce SCT from microorganisms.

2.2.3.1 Bacterial growth conditions

Escherichia coli K12 MG1655 (DSMZ 18039) was cultured in 1 mL of LB broth at 30°C and 750 rpm with the Thermomixer Comfort (Eppendorf, Germany) to exponential growth phase (OD600 of ~ 2.2-2.6). From this point forward, cells were processed in a UV decontaminated ISO 4 clean room. Equipment and gloves were decontaminated with DNA AWAY (Thermo Fisher Scientific, USA). Consumables were UV treated for 1 hr in a crosslinker and 1X PBS was UV treated for 6 hours in a 254nm shortwave ultraviolet crosslinker at 0.12 Joules/cm^2 (Analytik Jena US).

2.2.3.2 Cell isolation and lysis

From this point forward, all steps dealing with single, 10, and 100 cells were performed in a UV decontaminated ISO 4 cleanroom. Equipment and gloves were decontaminated with RNase AWAY (Thermo Fisher Scientific, USA). Consumables and 1X PBS were UV treated for 1 hr in a 254nm shortwave ultraviolet crosslinker at 0.12 Joules/cm^2 (Analytik Jena US).

Cells were first diluted to ~ 10^6 cells mL^{-1} with sterile 1X PBS. For the heat-shock cells, the 1X PBS was first heated to 50°C to prevent adverse transcriptional changes from temperature shock. A BD FACSMelody (Becton-Dickson, USA), fitted with a 100 µM nozzle, was used to sort cells into 384-well plates (Bio-Rad, USA) prefilled with 2.2 µL of 1X PBS + 0.1 U µL^{-1} RNasin® RI (Promega, USA). Plates were sealed with Microseal B (Bio-Rad, USA) and centrifuged at 4°C for 3 min and 3000 rpm (Eppendorf, Germany). The plates were then frozen in liquid nitrogen for 30 s and thawed at 21°C for 2 min. Lysed cells were immediately subjected to WTA.

2.2.3.3 WTA and library preparation

WTA was performed with reagents from the REPLI-g WTA Single Cell Kit (QIAGEN, Germany) but with several modifications to the protocol. All reaction volumes were reduced to 0.4x the volume of the original protocol (**Table 2.2**) and dispensed with the IDOT mini (Dispendix, Germany). Kit reagents, except for gDNA Wipeout Buffer, Quantiscript RT Enzyme Mix, Ligase Enzyme Mix, and Syto-13, were UV treated on an ice-water bath in a crosslinker for 30 min (Tanja Woyke et al., 2011). After all dispensing steps, the plate was sealed with Microseal B (Bio-Rad, USA) and briefly spun down for 10 s. Incubations were performed in a CFX384 Touch Real-Time PCR thermocycler (Bio-Rad, USA). In between incubation steps, the plate was kept on a cooling block to avoid degradation.

Table 2.2. Modified REPLI-g WTA Single Cell reaction volumes and components

Components	Reaction Volume (µL)	
Lysis		**2.20**
Phosphate Buffer Saline	2.19	
RNasin® RI (40 U/µL)	0.01	
gDNA Wipeout Buffer		**0.40**
Reverse Transcription		**1.40**
H_2O	0.10	
RT/Polymerase Buffer	0.80	
Random Primer	0.20	
RNasin® RI (40 U/µL)	0.10	
Quantiscript RT Enzyme Mix	0.20	
Ligation		**2.00**
Ligase Buffer	1.60	
Ligase Mix	0.40	
REPLI-g SensiPhi Amplification		**6.00**
REPLI-g sc Reaction Buffer	5.56	
REPLI-g SensiPhi DNA Polymerase	0.20	
SYTO-13 (50 µM)	0.24	
Total Final Volume		**12.00**

Following cell lysis, gDNA wipeout buffer was immediately added and cells incubated at 42°C for 10 min. During this time, the random primers were denatured at 65°C for 3 min, then immediately placed on ice for at least 1 min before being added to the RT mix. The oligo dT primers in the original RT mix were replaced with a final concentration of 4U/µL RI (Promega, USA) (**Table 2.2**). RT incubation included an extra, initial incubation for 10 min at 25°C before proceeding with RT which was extended from 60 min to 90 min at 42°C, followed by a deactivation step at 95°C for 3 min. Ligase mix was then immediately dispensed and incubated for 30 min at 24°C, followed by deactivation at 95°C for 3 min. The final REPLI-g SensiPhi amplification mix was modified to include SYTO-13 (Invitrogen, USA) at a final concentration of 2 µM (**Table 2.2**) to monitor exponential cDNA amplification with the CFX384 Touch Real-Time PCR thermocycler (Bio-Rad, USA). The amplification reaction incubation time was extended from 2 hrs to 12 hrs, followed by polymerase deactivation at 65°C for 5 min. The amplified cDNA was stored at -20°C until further processed.

The following steps were performed under a UV decontaminated Laminar Flow PCR workbench (STARLAB International GmbH, Germany) sterilized with DNA AWAY (Thermo Fisher Scientific, USA). Prior to library preparation, samples were cleaned-up using the DNA Clean & Concentrator-5 kit (Zymo Research, USA). cDNA concentrations were measured with Qubit™ DNA HS assay (Life Technologies, USA). All libraries were normalized to 12.32 ng and prepared using the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina® (New England Biolabs, USA) following the <100 ng protocol. Fragmentation was set to 14 min and 7 PCR cycles were used. NEBNext® Multiplex Oligos for Illumina® was used for barcoding. Library concentration and size was quantified with Qubit™ DNA HS assay (Life Technologies, USA) and Bioanalyzer High Sensitivity DNA kit (Agilent, USA). Pooled libraries were sequenced using an Illumina NextSeq 550 (2 x 75 bp paired-end) (Illumina, USA).

2.2.3.4 Bulk RNA extraction and library preparation

RNA from non-treated and heat-shocked cells was extracted using Direct-zol RNA Miniprep Kit (Zymo Research, USA) under a UV decontaminated Laminar Flow PCR workbench (STARLAB International GmbH, Germany) decontaminated with RNase AWAY (Thermo Fisher Scientific, USA). The recommended DNase I protocol from the kit was performed. An additional DNase treatment was applied after extraction using the TURBO DNA-free™ kit (Invitrogen, USA). RNA was quantified with the Qubit™ RNA HS kit and stored at -80°C until library preparation with 40U of RI (Promega, USA).

Bulk RNA libraries were prepared using the NEBNext® Ultra™ II RNA Library Prep Kit for Illumina® kit (NEB, USA). Input RNA was normalized to the same concentration. NEBNext® Multiplex Oligos for Illumina® were used for barcoding. Library concentration and size was quantified with Qubit™ DNA HS assay (Life Technologies, USA) and Bioanalyzer High Sensitivity

DNA kit (Agilent, USA). Pooled libraries were sequenced using an Illumina NextSeq 550 (2 x 75 bp paired-end) (Illumina, USA).

2.2.3.5 Data Processing

After demultiplexing, the raw read quality was assessed *via* FastQC v0.11.5 (www.bioinformatics.babraham.ac.uk/projects/fastqc). Cutadapt v3.5 (Martin, 2011) was run in paired-end mode to trim low-quality bases with a phred score lower than 30 from 3' ends and to remove Illumina adapters (AGATCGGAAGAGC). Read pairs were discarded if any of the reads were shorter than 20 nt. After trimming, read quality was assessed again. Potential contaminations were evaluated *via* FASTQ-Screen v0.15.2 (Wingett & Andrews, 2018) and the rRNA content was estimated using SortMeRNA v4.3.4 (Kopylova et al., 2012) (silva-bac-16s-id90 and silva-bac-23s-id98 databases for reference) with default settings. The trimmed reads were mapped to the *E. coli* MG1655 strain (ASM584v2, 28-11-2021) genome with the STAR aligner v2.7.6a (Dobin et al., 2013). Hereby, intron alignment was disabled. Read pairs were kept if the length-normalized alignment score and the length-normalized number of matched bases were at least 0.5. Chimeric junctions were reported if the mapped length of the segments was at least 20. RPM tracks were put out in bedGraph format and used for visualization with Integrative Genome Viewer (IGV) v2.11.9 (Thorvaldsdóttir et al., 2013). Read counts for each gene were determined *via* Feature Aggregate Depth Utility (FADU) v1.8 in paired-end mode and with default settings, enabling the counting of multimapping and overlapping reads (Chung et al., 2021).

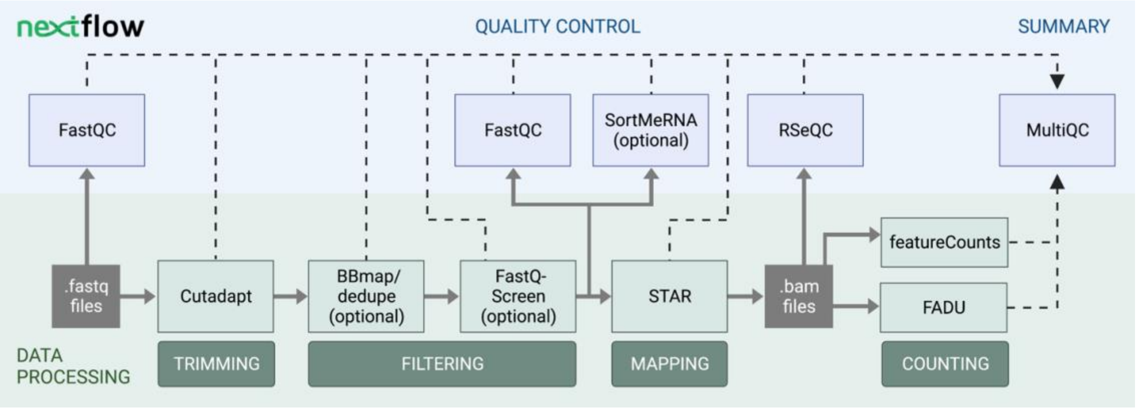


Figure 2.1. RNA-seq data analysis workflow with Nextflow
The workflow was implemented with Nextflow (DI Tommaso et al., 2017).

Finally, results were summed up with the MultiQC tool v1.11 (Ewels et al., 2016). The processing pipeline was implemented in Nextflow v21.04.3 (DI Tommaso et al., 2017) (**Figure 2.1)** and run on a LSF cluster system. Code will be made available at https://github.com/KIT-IBG-5.

2.2.3.6 Statistical Analysis

Counts were normalized and differentially transcribed genes were analyzed (padj = < 0.05) with DESeq2 (Love et al., 2014) *via* the statistical computing language R v4.0.0 (R Core Team, 2015). Ribosomal RNA genes were removed prior to analysis. Raw count data was transformed using variance stabilization transformation and subjected to PCA where the 300 most variable genes were taken into account. Normalized count values of the significant genes ($p<0.05$) were used for Gene Set Enrichment Analysis with the GSVA package (Hänzelmann et al., 2013). Bioconductor package org.EcK12.eg.db was used as annotation and only gene sets containing at least five genes were considered. In the case of pseudo-bulk samples, count matrices were generated by adding up raw counts of the respective single /multiple cell samples; analysis was performed as described above.

3 Results and Discussion

3.1 Chapter 1: Improving microbial single-cell genomics

There exist many species with unique and important ecological roles, such as nitrogen fixation, methane and methanol oxidation, respiratory dehalogenation, or secondary metabolite synthesis that can be found in low abundance in the environment (Dam & Häggblom, 2017; Griffiths et al., 2004; Pratscher et al., 2018). Obtaining genomes of such rare microorganisms has historically remained a large problem, with both culture-dependent and culture-independent techniques. The latter would require profound sequencing depths, which often becomes very expensive (Hugenholtz et al., 1998; Köpke et al., 2005) and therefore unfeasible. Unfortunately, this problem is also not fully resolved with standard SCG, not only due to the need for sorting tens of thousands of cells to access rare taxa, but also due to low success of cell lysis and biased high cost WGA (Kaster & Sobol, 2020). Therefore, the purpose of this thesis was to improve on the most crucial steps of the SCG workflow, namely labeling, sorting, lysis, and amplification, to help enrich ecologically important but low abundant microorganisms as well as increase the completeness of recovered SAGs for a better understanding of their metabolic potential.

3.1.1 A modified targeted-cell sorting method enriches minority taxa

FISH is a well-established method that uses fluorescently labeled oligonucleotide probes for taxon-specific labeling of prokaryotes (Amann et al., 1990). Standard FISH protocols are performed on glass slides and use fixatives, such as paraformaldehyde, to maintain cell wall integrity, as well as allow probes to more easily penetrate the cell and increase fluorescence signal intensity (Shakoori, 2017). However, fixatives create cross links between nucleic acids, rendering the cells useless for genomic applications (Clingenpeel et al., 2014; Doud & Woyke, 2017). Therefore, a modified targeted-cell sorting method for SCG was established here which removes the need for a glass-side and fixatives, referred to as in-solution, fixation-free FISH (Haroon et al., 2013; Podar et al., 2007; Yilmaz et al., 2010). This method was used to enable the labeling of minority member taxa of interest from an environmental community. Specifically, we chose to survey microbial communities within winery waste water treatment plant (WWTP) effluent samples collected in 2013, 2014, and 2015 from the Juanicó winery in Canelones, Uruguay (Dam et al., 2020). Metagenomics was used to assess the total microbial communities within all samples, which showed that the three samples highly differed in relative abundances of almost all phyla (**Figure 3.1**). It is likely that these differences in abundances are due to large fluctuations of organic and inorganic effluent into the WWTP, which changes the physiochemical environment and in turn impacts the microbial community (Mosse et al., 2012). Notably, one of the most striking differences in relative abundances was seen in the bacterial phyla Chloroflexi by a factor of 10 (**Figure 3.1)**. Hence, Chloroflexi was a good choice as the taxa of interest for targeted-labeling in this proof of principle study.

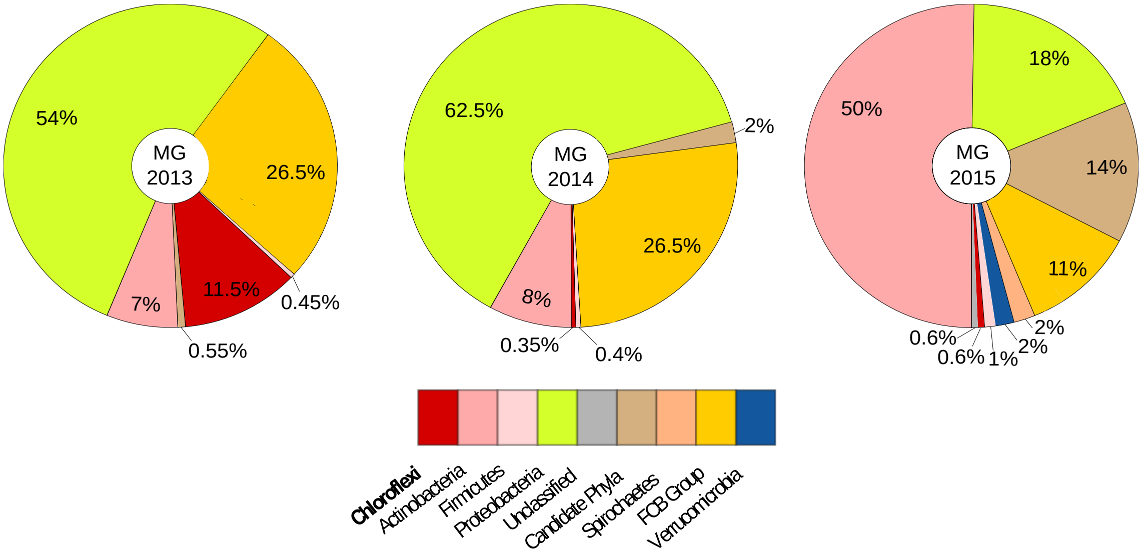


Figure 3.1. Taxonomic overview of Uruguayan WWTP samples
Metagenomes were sampled from the same WWTP in 2013, 2014, and 2015. Classification and relative abundances were determined from MAGs. Relative proportions between MAGs were determined based on average contig coverages of each MAG in relation to the overall coverage of binned contigs from the metagenome. The resulting relative abundances between binned taxa largely reflect the corresponding taxon proportions of the community observed *via* marker gene analyses of the complete metagenomes. MG = metagenome. Published in Dam et al. (2020).

Metagenomic and 16S rRNA data show that Chloroflexi are ubiquitous throughout the environment, however, they have also mostly evaded cultivation attempts and are therefore under-characterized, although they have potential for important biotechnological applications (Kaster et al., 2014; Kogawa et al., 2022; Loffler et al., 2013; Zheng et al., 2019). As of July 2020, approximately only 15% of Chloroflexi genomes are represented by a cultured isolate, according to Genomes OnLine Database (GOLD) (Mukherjee et al., 2019) (**Figure 1.1**). Previously, this phyla consisted of eight classes, but because of the massive increase in culture-independent efforts today, the phyla has expanded to 12 classes, most of which remain unresolved, according to Genome Taxonomy DataBase (GTDB) (Parks et al., 2022).

In order to obtain sufficient fluorescence signal from low abundant Chloroflexi, longer hybridization times and higher probe concentrations were used to overcome the problem of low fluorescent signals since no fixative was used. Targeted-cell sorting was first validated using a mock mixed microbial culture containing 1% of a known Chloroflexi isolate, *Sphaerobacter thermophilus* (DSM20745) and 99% *Escherichia coli* K12 (DSM498) before testing on the WWTP sample. The cell-labeled mixture was sorted in two consecutive steps: first an enrichment sort and then a single cell sort. The sorted population was gated based on its greater fluorescent signal compared to the non-labelled mixed culture. The gated population was greatly enriched from 0.9% to 76% after the first sorting step (**Figure 3.2**). Next, the modified FISH approach was applied to the 2015 WWTP sample which contained approximately 0.6% Chloroflexi and sorted with FACS (**Figure 3.3**). The gated population showed much greater fluorescence signal than the non-labeled control sample under both epifluorescence microscopy and with FACS (**Figure 3.3**). The sorted cells were then subjected to WGA with MDA.

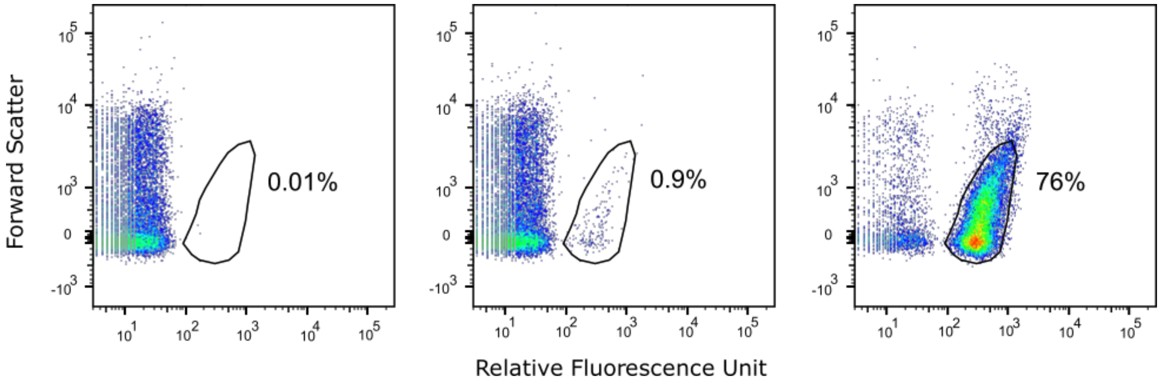


Figure 3.2. FACS analysis of in solution, fixation-free FISH applied to a mock community
A mock community of 1% *S. thermohilus* and 99% *E. coli* was treated with the in solution, fixation-free FISH protocol and sorted with FACS. **A** Negative control without probes. **B** Gated population after the first sort. **C** Gated population after the second sort. Cells in the mixed culture were labeled with two Chloroflexi-specific probes (CFX1223 and GNSB941). A total of 20,000 events were recorded. Numbers indicate the percentage of the gated population in the total event recorded.

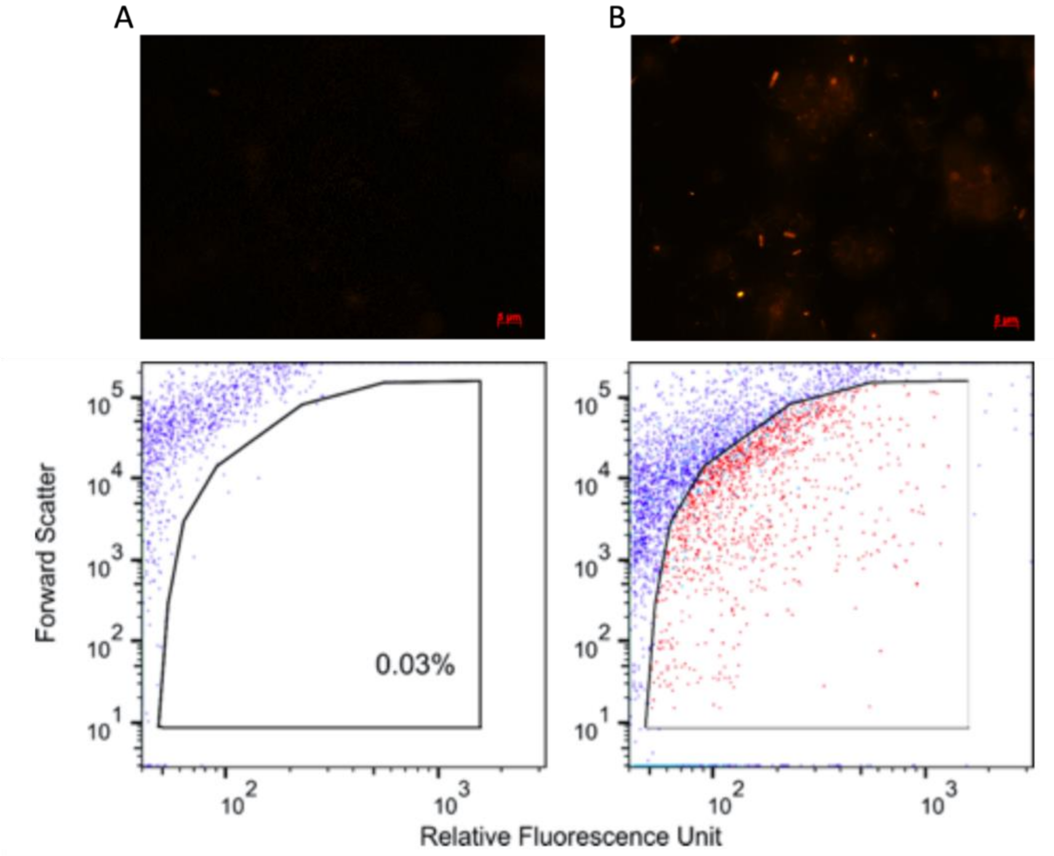


Figure 3.3. Epifluorescence microscopic and FACS of targeted Chloroflexi cells
A Negative control, without Chloroflexi FISH probes. **B** FISH-labeled Chloroflexi cells from the winery WWTP 2015 sample. Relative fluorescence units (RFU) refer to the fluorescent signal of SYBR™Green, which enters the cell membrane and binds to dsDNA and is used to distinguish the cells from any non-living material.

Following MDA, the success of amplified SAGs was on average 38.6%, which is within those reported from studies using conventional SCGs (Kaster et al., 2014; Rinke et al., 2014; Swan et al., 2011). The average SAG completeness was only 32%, likely due to the biased nature of MDA against high GC genomes (Sabina & Leamon, 2015) typically found in Chloroflexi (Kaster et al., 2014). However, utilizing WGA-X™ with a more thermotolerant phi29 polymerase (**Table 1.1**) instead of MDA, could result in better recovery of high GC organisms (Stepanauskas et al., 2017). Additionally, ethanol which was used here as a pre-treatment to increase probe penetration, has been reported to reduce genome coverage possibly due to ethanol causing protein aggregation around the DNA, preventing the polymerase from functioning (Clingenpeel et al., 2014).

Fortunately, it seems likely that this step can be removed in the future as it may not be entirely necessary for probe penetration (Yilmaz et al., 2010), but this needs to be confirmed with more diverse cell types. Lastly, to further overcome the problem of cell recovery and genome completeness, other improvements to the cell isolation, cell lysis and WGA steps established in this dissertation should be used in the future.

3.1.2 Targeted-cell sorting captures rare taxa and their metabolic potential

In order to prove the targeted-cell sorting's ability to enrich capturing of minority member's genomes from an environmental sample, the recovered Chloroflexi SAGs were compared to the Chloroflexi MAGs from the 2015 WWTP sample. However, only two MAGs could be retrieved from the 2015 sample alone, therefore differential coverage binning was performed on all three samples. This only resulted in four Chloroflexi MAGs, which were classified as *Ardenticatenia*, *Thermomicrobia*, and *Anaerolinea* (**Table 3.1**). From the 38.6% amplified SAGs, 41 could be identified as Chloroflexi with 16S rRNA screening. After assembly, genomes which shared greater than 99% identity as determined by 16S rRNA screening and more than 98% average nucleotide identity (ANI) over at least 10% of the genome, were co-assembled into one genome (CAGs). CAG-1 was the result of 7 combined SAGs, CAG2 from 4 SAGs, and CAG3 from 4 SAGs (**Table 3.1**). This left a remaining 19 SAGs of Chloroflexi, of which nine were classified as *Anaerolineae*, six as *Caldilineae*, three to *Ardenticatenia*, and one to *Chloroflexia* (**Table 3.1**). Both SAG7 and SAG16 could not be placed into the defined classes and were therefore determined as unclassified and within candidate class *Thermofonsia*, respectively.

Table 3.1. Overview of Chloroflexi CAGs and SAGs collected with targeted SCG

Genomes	Completeness* (%)	Adj. contamination*	Size (Mbp)	GC (%)	Classification
CAGs and SAGs					
CAG1	89.81	8.7	6.97	53.4	*Caldilineae*
CAG2	75.71	2.3	3.41	62.5	*Anaerolineae*
CAG3	28.74	0	1.91	58.9	*Caldilineae*
SAG4	55.17	1.72	2.12	61.3	*Anaerolineae*
SAG5	49.14	0	3.37	52.9	*Caldilineae*
SAG6	49.06	1.72	1.73	49.1	*Anaerolineae*
SAG7	45.05	0.60	0.63	36.1	Unclassified
SAG8	23.82	0.16	0.72	50.6	*Ardenticatenia*
SAG9	23.67	0	1.42	63.0	*Caldilineae*
SAG10	20.06	0.34	1.10	57.6	*Ardenticatenia*
SAG11	19.28	0	0.51	60.6	*Caldilineae*
SAG12	18.97	0	0.72	50.6	*Anaerolineae*
SAG13	16.85	0	0.88	47.3	*Anaerolineae*
SAG14	16.14	1.72	0.58	45.4	*Anaerolineae*
SAG15	14.42	0.16	1.18	61.9	*Ardenticatenia*
SAG16	14.33	0	1.72	59.0	Cand. *Thermofonsia*
SAG17	13.95	0	0.87	45.4	*Anaerolineae*
SAG18	10.82	0	0.54	60.0	*Chloroflexia*
SAG19	5.17	0	0.12	46.6	*Caldilineae*
MAGs					
MAG1	92.37	3.89	5.32	61.4	*Ardenticatenia*
MAG2	67.63	0	1.69	60.5	*Thermomicrobia*
MAG3	75.24	0.67	1.64	55.7	*Anaerolineae*
MAG4	48.12	1.72	3.09	64.4	*Ardenticatenia*

[a] CAGs were co-assembled from multiple SAGs as followed when 16S rRNA gene sequences resulted from screening shared at least 99% similarity and the average nucleotide identity values (ANI) determined by the Pyani package of preliminary bins shared more than 98% over at least 10% genome coverage.

[*] Completeness and contamination estimations were based on CheckM (Parks et al., 2015) results using bacterial-specific marker sets. The CheckM "contamination" estimate, which was based on the number of duplicate markers may not reflect actual contamination, was adjusted by excluding duplicates with near identical gene sequence (the original CheckM "strain heterogeneity" estimate).

This revealed that the targeted SCG approach retrieved 10x more Chloroflexi genomes than metagenomics from the 2015 winery WWTP sample alone and 5x more genomes compared to the differential coverage binning of the 2013, 2014, and 2015 samples. MAG2 was the only genome not captured with SCG, likely because it belongs within the family *Sphaerobacteriaceae* which are notoriously hard to lyse due to their tough cell wall (Pati et al., 2010). This suggests that the lysis used here was not sufficient to lyse this species, and highlights the need for further optimization of the cell lysis step in the future to increase the recovery of hard to lyse cells. Among the rest of the MAGs, MAG1 was found to be retrieved by targeted cell sorting as determined by a 99% ANI over 11–17% genome coverage with SAG8, SAG10, and SAG15. Therefore, it is probable that these genomes originate from the same *Ardenticatenia* species. Even though MAG1 was on average 73% more complete than its related SAGs, it appeared that MAG1 was largely missing putative genomic islands and mobile genetic elements by direct comparison to the SAGs. This included a 54.82 kb putative phage contig in SAG9 and several instances of transposon associated genes of various putative functions in the other SAGs. This confirms a known drawback of metagenomic binning algorithms, which find it difficult to assemble these structural and mobile genetic elements since they are typically different from the rest of the genome (Maguire et al., 2020). Additionally, MAG1 had a higher estimated percent contamination compared to the SAGs. This is likely due again to the binning algorithms, which have an increased chance to assemble multiple genomes of the same species together if their genomes are not too genetically different (i.e. strain variants) (Dick et al., 2009).

Even though the overall genome recovery and completeness of the SAGs was lower than that of the MAGs, a look into potential metabolic functions of the SAGs found that the Chloroflexi in the winery WWTP are heterotrophic and include genes involved in the transformation and degradation of carbohydrates and aromatic compounds. SAG11, which clustered within *Caldilineae* and was only retrieved *via* the targeted SCG approach, contained marker genes indicative of carbon fixation *via* the Calvin–Benson–Bassham (CBB) cycle for the first time, which was recently independently confirmed (West-Roberts et al., 2021). Additionally, our finding of a potential secondary metabolite producer in CAG1, belonging to the class *Caldilineae*, is in agreement that this class might become an emerging candidate for production of biologically active compounds (Bayer et al., 2018; Kogawa et al., 2022), highlighting the feasibility of this approach to target interesting and biotechnologically important organisms.

Overall, we could unlock novel phylogenies, metabolisms, and other physiological characteristics of rare members of the community that would have otherwise been overlooked by conventional metagenomics. This is especially illuminated by the identification of several potential genomic islands related to horizontal gene transfer in the SAGs but not found within the corresponding MAG, as such regions are unlikely to be correctly and unambiguously binned from metagenomes. Moreover, by placing a focus on specific organisms of interest, targeted-cell sorting helps to reduce the costs of SCG by reducing the number of cells needed to isolate low abundant taxa. Hence, this technique represents an essential complement to metagenomics and microbial community-focused research approaches for elucidating the genomic potential of novel taxa currently still hidden within microbial dark matter.

3.1.3 Cell printers cause less damage to FISH-treated cells

FACS has remained the standard sorting method for most SCG projects due to its high-throughput sorting, that it can typically be equipped with 3 or more lasers to target fluorescently labelled cells, and that it is commercially available and therefore easily accessible (Blainey, 2013). However, one major issue with FACS is that it operates at a high pressure (~80 psi), which could lead to premature cell lysis (Blainey, 2013; Mollet et al., 2008; Wiegand et al., 2021) for easy to lyse cells (i.e. Gram-negatives), cells who already have damaged cell walls prior to sorting (i.e. FISH-treated cells) (Clingenpeel et al., 2014), and/or anaerobic cells who will lyse upon exposure to oxygen. While in theory we need cells to lyse for successful WGA, premature lysis makes the DNA more susceptible to degradation after sorting and during the subsequent lysis step, which leads to lower genome recovery (Rinke et al., 2014). Also, this extracellular DNA can cause contamination in other SAGs if sorted with the cells. To help mitigate these issues, cell printing technology which uses inkjet printer heads to more gently deposit cells (Gross et al., 2013; Riba et al., 2016) could be used.

In order to understand the extent of damage during FACS to pre-damaged cells with different cell walls, comparisons between the viability of FISH-treated, Gram-negative

Flavobacterium denitrificans and Gram-positive *Bacillus marisflavi* after FACS and cell printing, were made here. Triplicates of 10 and 100 cells were sorted for each and the effect of FISH and FISH without probe was assessed five days after growth. FISH treatment with probe was seemingly more damaging to cells than without, likely because the act of taking up large sizes of DNA causes additional stress to the cell (Siguret et al., 1994) (**Table 3.2**). Furthermore, we could infer that FACS causes more damage to pre-damaged cells since there was less visible growth than with the cell printer (**Table 3.2**). This effect was more prominent in the Gram-negative bacteria who are more susceptible to damaged cell walls. Recently, a study that modified FISH for the sole purpose of keeping cells alive after FACS sorting, was still only able to grow approximately 6% and 3% of Gram positive and Gram negative cells, respectively (Batani et al., 2019). In the future, it would be interesting to see how much more viable cells could be grown with their modified FISH protocol subjected to cell printing, since they did not assess the effect of FACS on their bacteria. Additionally, the difference in SAG recovery and completeness following FISH treatment still needs to be assessed with the cell printer, as this has only been done previously with FACS (Clingenpeel et al., 2014). However, recent results do provide evidence that the cell printer increases the average successful SAGs from a given experiment up to 80% when applied to non-FISH treated cells (Wiegand et al., 2021). This further highlights the promising future applications for this cell isolation method.

Table 3.2. Viability after cell sorting with FACS and B.SIGHT™

Organism	Gram Stain	Probe	FACS		Cell Printer	
			10	10^2	10	10^2
Flavobacterium denitrificans	-	no	-	-/+	-	+
		yes	-	-	-	+/-
Bacillus marisflavi	+	no	-	-/+	+/-	+
		yes	-	-/+	-	+/-

-, no growth; -/+, growth in 1 or 2 out of 3 replicates; +, growth in all replicates

3.1.4 Improved cell lysis provides better genome recovery

Cell lysis is one of the most important but also most challenging steps of SCG because it needs to be capable of lysing different types of microbial cell walls without damaging the DNA. Also, as discussed above, if cells are prematurely lysed then the DNA will likely be damaged when lysis buffer is applied. DNA damage is a problem for the standard WGA method MDA, because it uses phi29 which requires larger DNA templates to begin synthesis (Dean et al., 2001), thus DNA damage largely contributes to lower SAG recovery and completeness (Clingenpeel et al., 2014). Due to these issues, each sample type and its treatment should be taken into consideration to optimize the cell lysis step. Here, the lysis was optimized for Gram-negative *E. coli*, since the subsequent experiments in this dissertation used this bacterium. In this work, the alkaline cell lysis buffer published in Stepanauskas et al. (2017) for WGA-X™ was chosen for modification since the cell lysis supplied in the REPLI-g Single Cell kit is proprietary. The original WGA-X ™ cell lysis buffer consists of 0.4 M KOH, 10 mM EDTA and 100 mM, and is neutralized with 1M Tris-HCl (Stepanauskas et al., 2017). Single *E. coli* cells sorted into WGA-X lysis with FACS and treated with one freeze-thaw cycle resulted in no cell amplification as it was likely to harsh for *E. coli* (**Figure 3.4**). The WGA-X lysis buffer was also diluted to 50% so that the final concentrations of the components were 0.2 M KOH, 5 mM EDTA, and 50 mM DTT. This resulted in 5 out of 6 replicate single cells amplifying (in green) (**Figure 3.4**). As a control, cells (in blue) were not provided with lysis buffer and only subjected to the one freeze-thaw cycle. Only three out of six control cells amplified with generally lower relative fluorescence intensities and higher cycle quantification (Cq) values, indicative of less DNA amplified likely due to inefficient denaturation. As the 50% WGA-X lysis buffer worked better, it was used for subsequent SCG tests in the following sections.

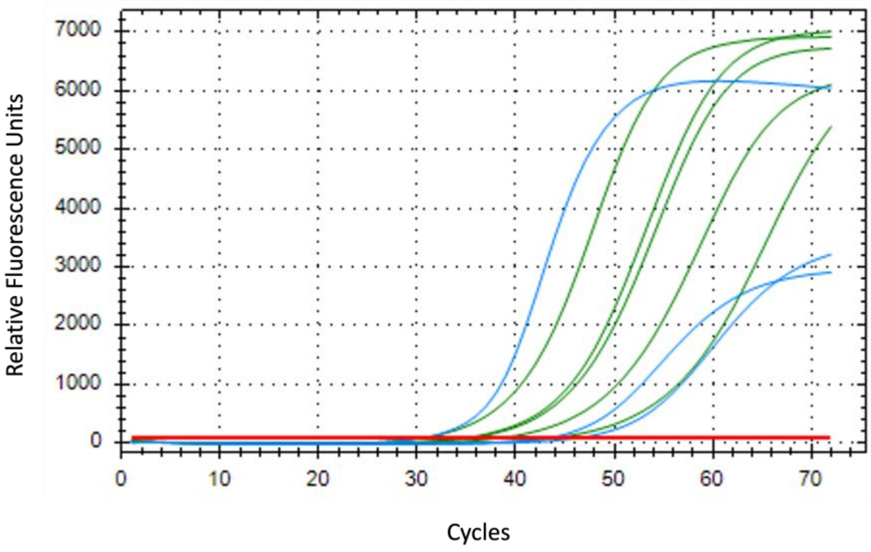


Figure 3.4. Comparing different single-cell lysis conditions
50% lysis in green, no lysis in blue, negative control in red. 100% WGA-X lysis resulted in no amplification and is therefore not shown. Each cycle represents 5 minutes of amplification time. Relative fluorescence units (RFU) refer to the fluorescent signal of SYTO™-13 measured with a real-time thermo-cycler. SYTO™-13 is used to monitor the progression of MDA because it binds to double-stranded DNA as it is amplified.

3.1.5 MDA volume reduction in multi-well plates improves genome coverage and uniformity

Due to amplification bias caused by MDA, the analysis of single-cell genomic data has remained challenging. This bias largely stems from random mechanisms, bias against high GC DNA and secondary structures as well as, exogenous and/or endogenous contamination (Lasken & Stockwell, 2007; Sabina & Leamon, 2015). Previous studies have observed that by simply reducing the total reaction volume of MDA and other WGA methods, amplification bias can be lessened (De Bourcy et al., 2014; Gole et al., 2013; Leung et al., 2016; Marcy, Ishoey, et al., 2007; Nishikawa et al., 2015; Rhee et al., 2016; Ruan et al., 2020). Furthermore, the costs for classic MDA in standard 50 µL reaction is ~24 USD (**Table 1.1**), which quickly becomes unaffordable and therefore, not applicable to most studies. Hence, volume reduction would also help to make SCG more accessible and more high-throughput.

Studies show that volume reduction improves polymerase specificity through "molecular crowding" (Minton, 2001; Zimmerman & Harrison, 1987). Molecular crowding reduces competition between amplification of the template and contamination by increasing the probability that polymerase and primers bind to template DNA and reducing spurious binding (Minton, 2001; Zimmerman & Harrison, 1987). Moreover, lower reaction volumes reduce the amount of surface area for nonspecific adsorption of nucleic acids to the multi-well plate walls (Belotserkovskii et al., 1996; Gaillard & Strauss, 1998). However, too much crowding can also cause adverse effects by reducing the polymerase from accessing the template (Kuznetsova et al., 2014; Ralston, 1990). To test this theory, we sorted single *E. coli* cells into 384-well plates to compare SAG amplification bias within total MDA microliter and sub-microliter reactions for the first time. In contrast, previous studies have largely examined volume reduction in the sub-nL to pL range (De Bourcy et al., 2014; Goldstein et al., 2017; Gole et al., 2013; Leung et al., 2016; Marcy, Ouverney, et al., 2007; Nishikawa et al., 2015; Rhee et al., 2016; Ruan et al., 2020).

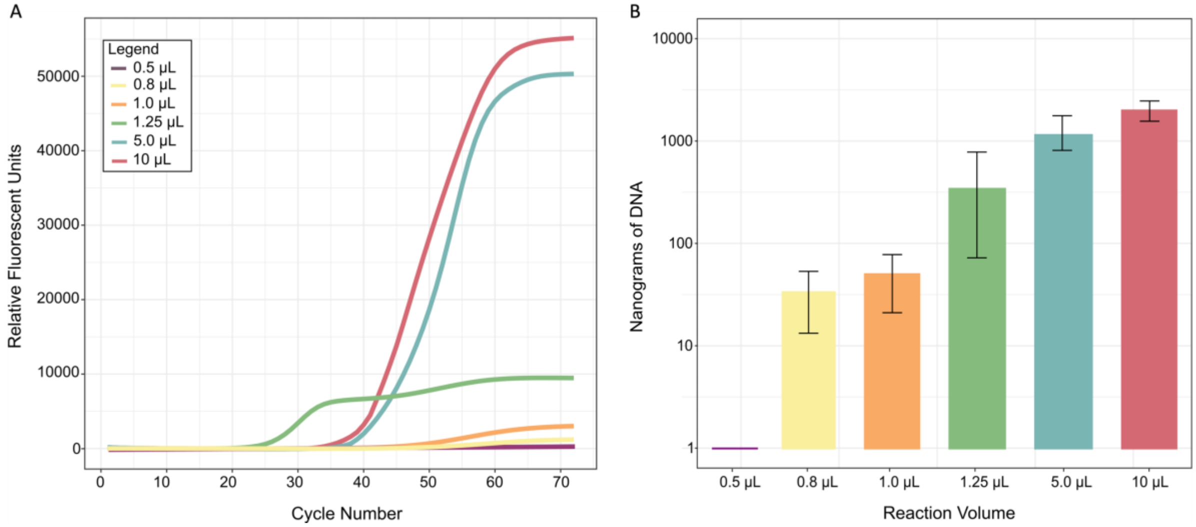


Figure 3.5. General MDA statistics overview
A Average MDA reaction kinetics by reaction size. **B** Average MDA amplification yield by reaction size. Relative fluorescence units (RFU) refer to the fluorescent signal of SYTO™-13 measured with a real-time thermo-cycler. SYTO™-13 is used to monitor the progression of MDA because it binds to double-stranded DNA as it is amplified. Standard error bars represent the standard deviation calculated using all five replicates from each reaction volume.

MDA's with total reaction volumes of 0.5 µL, 0.8 µL, 1.0 µL, 1.25 µL, 5.0 µL, and 10 µL were conducted in 384-well plates. The lowest-sized MDA reaction, 0.5 µL, did not work, likely due to evaporation and/or sterical hinderance of the polymerase in small volumes (Kuznetsova et al., 2014; Ralston, 1990). The average detected amplification progress and DNA yield from the successful reactions increased with reaction volume (**Figure 3.5A-B**). The lower amplification gain and DNA yield in the smaller reactions initially indicated that volume reduction likely limited the exponential nature of MDA (De Bourcy et al., 2014; Rhee et al., 2016), which should improve genome coverage and uniformity. To further compare the quality of the WGA reactions, a total of five successfully amplified replicates from all volumes were subjected to Illumina sequencing using equal amounts of DNA (**Appendix Table 1**).

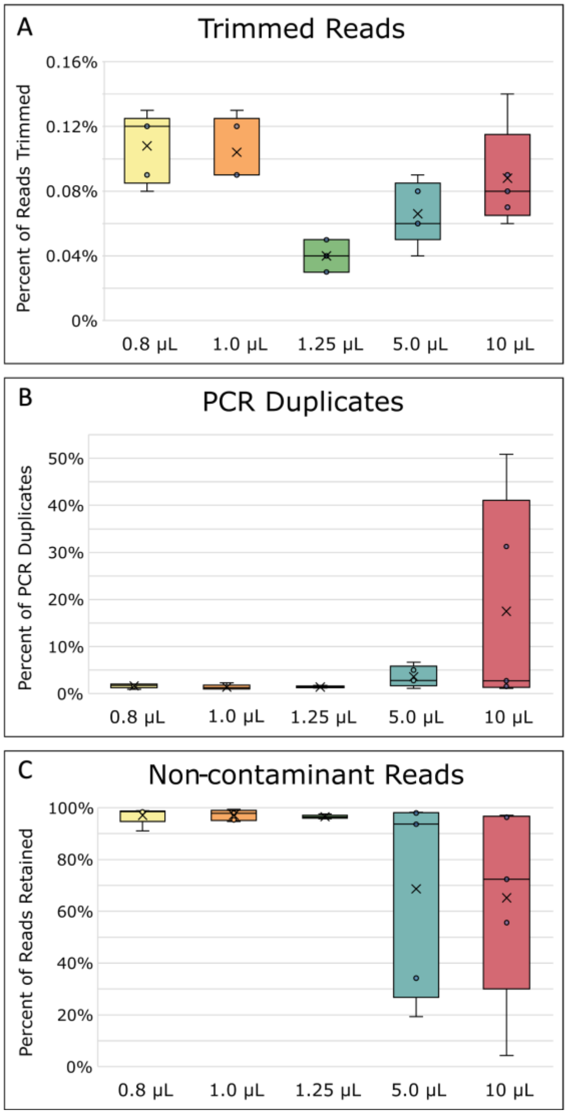


Figure 3.6. Read processing statistics

A Average percentage of reads removed during quality trimming. **B** Percentage of PCR duplicates removed. **C** Average percentage of reads kept after read contaminant filtering. The boxes middle line represents the median, and the x represents the mean. Five replicates were used for calculation.

On average, the MDA reactions performed in 1.25 µL volumes lost significantly fewer reads to read quality trimming (**Figure 3.6A**) compared to all other reaction volumes (p=0.0002, **Appendix Table 2**). After read trimming, all samples were normalized to 200X sequencing depth before further read processing steps. After depth normalization, the number of duplicated reads was, on average, greater in larger sized volume reactions (**Figure 3.6B**), but the difference between all reactions of the different volumes was not found to be significant (p=0.0870, **Appendix Table 2**). While some amount of read duplication inevitably results from MDA's exponential amplification nature, comparisons of the percent duplicates between samples could still provide insight into the specificity of the amplification itself. A higher number of duplicates can be caused by the lower template specificity in large MDA reactions causing more spurious priming and amplification (Leung et al., 2016; Marcy, Ishoey, et al., 2007), especially when template concentrations are very low (Bansal, 2017). Furthermore, the issue of lower template specificity also explains why there was an observed trend that larger reaction volumes had more contaminant reads removed after filtering than the smaller reactions (**Figure 3.6C**). Lower specificity, leading to more contamination, is likely due to the increased competition between background contamination and the *E. coli* single-cell DNA (Gole et al., 2013; Marcy, Ishoey, et al., 2007). This increase in contamination was also reflected in the higher amplification gain and product yield mentioned previously (**Figure 3.5A-B**), which other studies reported as well (Gole et al., 2013; Nishikawa et al., 2015; Rhee et al., 2016). In general, we also observed that 5 µL and 10 µL MDA reaction volumes gave less consistent results, as evidenced by larger variation between replicates (**Figure 3.6**).

As a consequence of lower template specificity, the MDA reaction volumes above 1.25 µL also performed worse during read mapping to the reference *E. coli* MG1655 genome, as indicated by genome coverage breadth and coverage uniformity (**Figure 3.7A-B**). MDA in 0.8 µL and 1.0 µL reaction volumes also resulted in low coverage breadth and uniformity, and a reaction volume of 1.25 µL was therefore determined as the "sweet-spot" for improved MDA in 384-well plates. Likely, the 0.8 µL and 1.0 µL reaction volumes were simply too low, causing too much molecular crowding, sterically hindering the polymerase from fully accessing the template DNA (Kuznetsova et al., 2014; Ralston, 1990), and/or there was too much evaporation. Reduced genome coverage was also recently reported for MDA reaction volumes below 150 nL on a microfluidic system (Ruan et al., 2020), suggesting that platforms independently have a specified “sweet-spot” for efficient MDA.

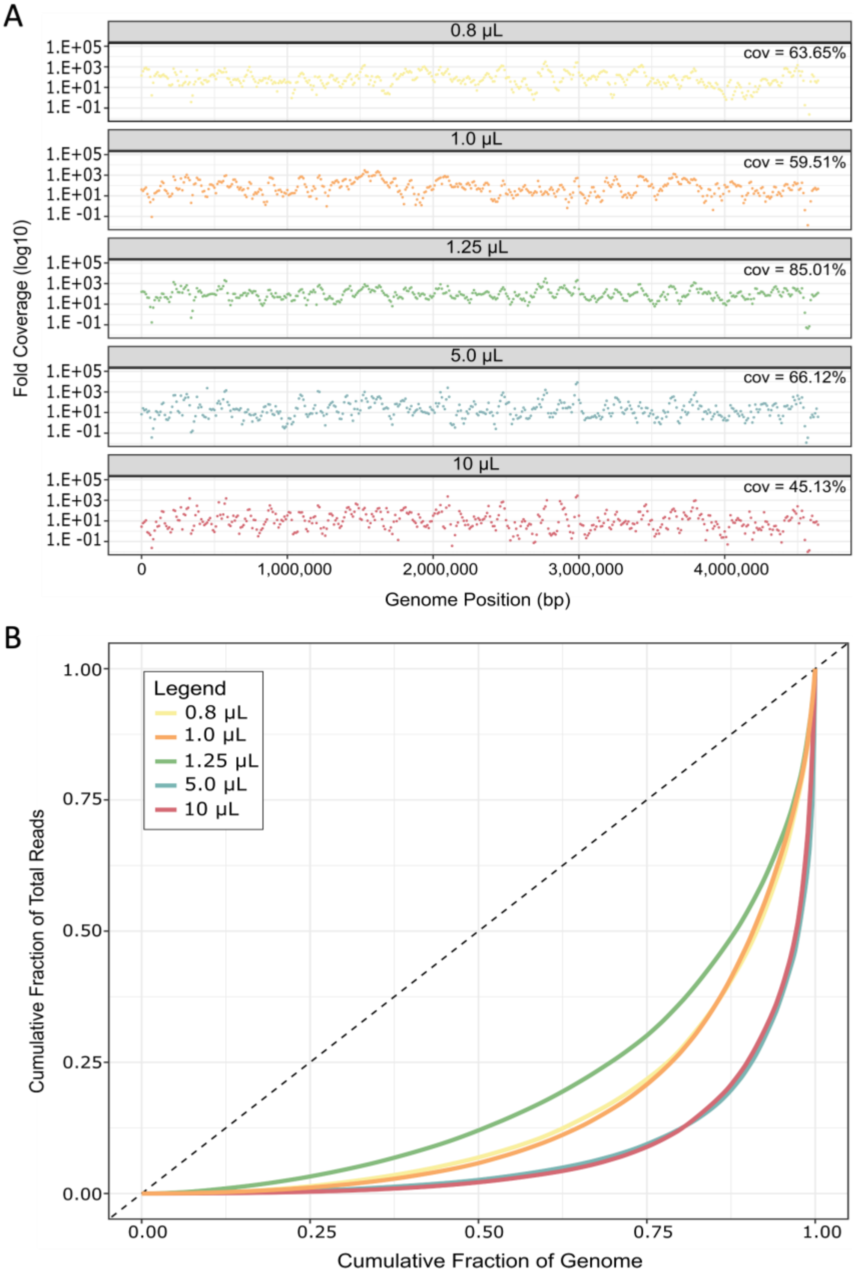


Figure 3.7. Genome coverage and uniformity bias

A Read depths from each replicate were calculated in 10 kb bins across the E. coli genome. Plots show average across all replicates for each reaction volume. Cov. is the average coverage breadth, i.e. the percentage of genome positions covered by at least one read. **B** Uniformity of read coverage and depth were calculated across 10 kb bins along the E. coli genome and averaged for all five replicates of each reaction volume.

On average, reads from 1.25 µL MDA reaction volumes covered 85± 13% of the *E. coli* genome, which was 19 to 40% more than the other sized reactions (**Figure 3.7A**). This increase in coverage was a large improvement when compared to current, well established methods like WGA-X™, which gives a reported ~36 ± 21% read coverage of *E. coli* in a standard 10 µL reaction (Stepanauskas et al., 2017). When compared to 10 µL reactions in this study, we still noted approximately 19% greater coverage breadth than WGA-X™, even though we used ~2 million fewer reads during read mapping. Likely, this difference can be attributed to the lysis modified specifically for *E. coli* herein. Furthermore, the average genome coverage in our study is ~45% greater than MDA performed in ~60 nL hydrogel reactions (Xu et al., 2016). Here, the much lower coverage for *E. coli* could be due to the fact that the authors performed a second-round of MDA, which has been shown to increase bias (De Bourcy et al., 2014). Our reported coverages are also well within range of those reported from a different nanoliter microfluidic method (De Bourcy et al., 2014), as well as from picoliter droplet reactions (Nishikawa et al., 2015), at the same sequencing depth (**Appendix Figure 1**). It should be mentioned that one other study reports ~15% greater coverage from MDA in nanoliter microwells when compared to our 1.25 µL average genome coverage at the same sequence depth (20X) (**Appendix Figure 1**) (Gole et al., 2013), however, the authors only used three single *E. coli* cells for testing.

To assess the uniformity of read coverage across the genome, reads were averaged into 10 kilo-base (kb) bins and their read depths plotted to visualize coverage depth for each reaction volume (**Figure 3.7A**). Especially in the larger volumes, more genome regions are not covered by any reads in comparison to MDA performed in 1.25 µL volumes. Furthermore, coverage depths were much more uniform across the genome in 1.25 µL, as evidenced by the smoother Lorenz curves (Zeileis et al., 2014) (**Figure 3.7B**). We further verified this by calculating the Gini index of each sample, which is a measure of deviation from uniformity ranging from 0 (perfectly uniform distribution) to 1 (extremely uneven distribution) (Dorfman, 1979). The Gini index differs significantly between different reaction volumes (p=0.0009, **Appendix Table 2**), and is lowest for 1.25 µL reactions (~0.7± 10.07, **Appendix Table 2**). These levels of uniformity are similar to those obtained from *E. coli* in 150 nL microfluidic MDA reaction volumes (De Bourcy et al., 2014) and in hydrogels (Xu et al., 2016).

Next, we assembled and compared SAGs for all replicates. Prior to assembly, read depths were normalized due to the large differences introduced *via* MDA, setting a target depth of 100X. However, MDA reaction volumes less than and greater than 1.25 µL resulted in lower final sequence depths due to the fact that more reads were lost during the read pre-processing steps (**Figure 3.8A**). Therefore, the resulting assemblies were of lower quality compared to assemblies from 1.25 µL MDA reaction volumes (**Figure 3.8B-C**). Specifically, 1.25 µL reactions had the longest average total length and N50 at 3,522,851 bp and 46,179 bp, respectively (**Figure 3.8B-C)**. N50 constitutes the minimum contig length above and below which 50% of the assembly's sequence information is contained and indicates that assemblies from 1.25 µL reaction volumes were more contiguous, resulting in higher quality assemblies than the other MDAs. Next, assembly coverage and completeness were calculated. The difference between these two measurements is that coverage is calculated as the percentage of the assembly (contigs) mapped to the reference genome (Gurevich et al., 2013), whereas genome completeness was estimated by MDMcleaner (Vollmers et al., 2022). In general, the assembly coverage ($p=0.009$) and completeness ($p=0.0128$) both significantly differed between the different sized reactions, based on one-way ANOVA (**Appendix Table 3**). Not surprisingly, coverage and completeness were highest for assemblies from 1.25 µL MDA reactions and were on average ~75± 14% and 94± 0.04%, respectively, while contamination was lowest (**Figure 3.8D-F**). Three out of five 1.25 µL MDA reaction replicates even achieved over 75% coverage, with the highest being 89.5% (**Appendix Table 3**). Comparatively, WGA-X™ reported *E. coli* assembly coverages of <60%, even with ~5X more reads (Stepanauskas et al., 2017). Whereas at 10 µL, our assembly coverages were found to be within the range of those reported from WGA-X™ in 10 µL reactions, highlighting how WGA-X™ could also benefit from further volume reduction. In comparison to other volume reduction approaches, our higher assembly coverages were within range of previously reported *E. coli* MDA coverages in pL droplets (88-91%) (Nishikawa et al., 2015) and nL wells (88-94%) (Gole et al., 2013) at similar sequence depths.

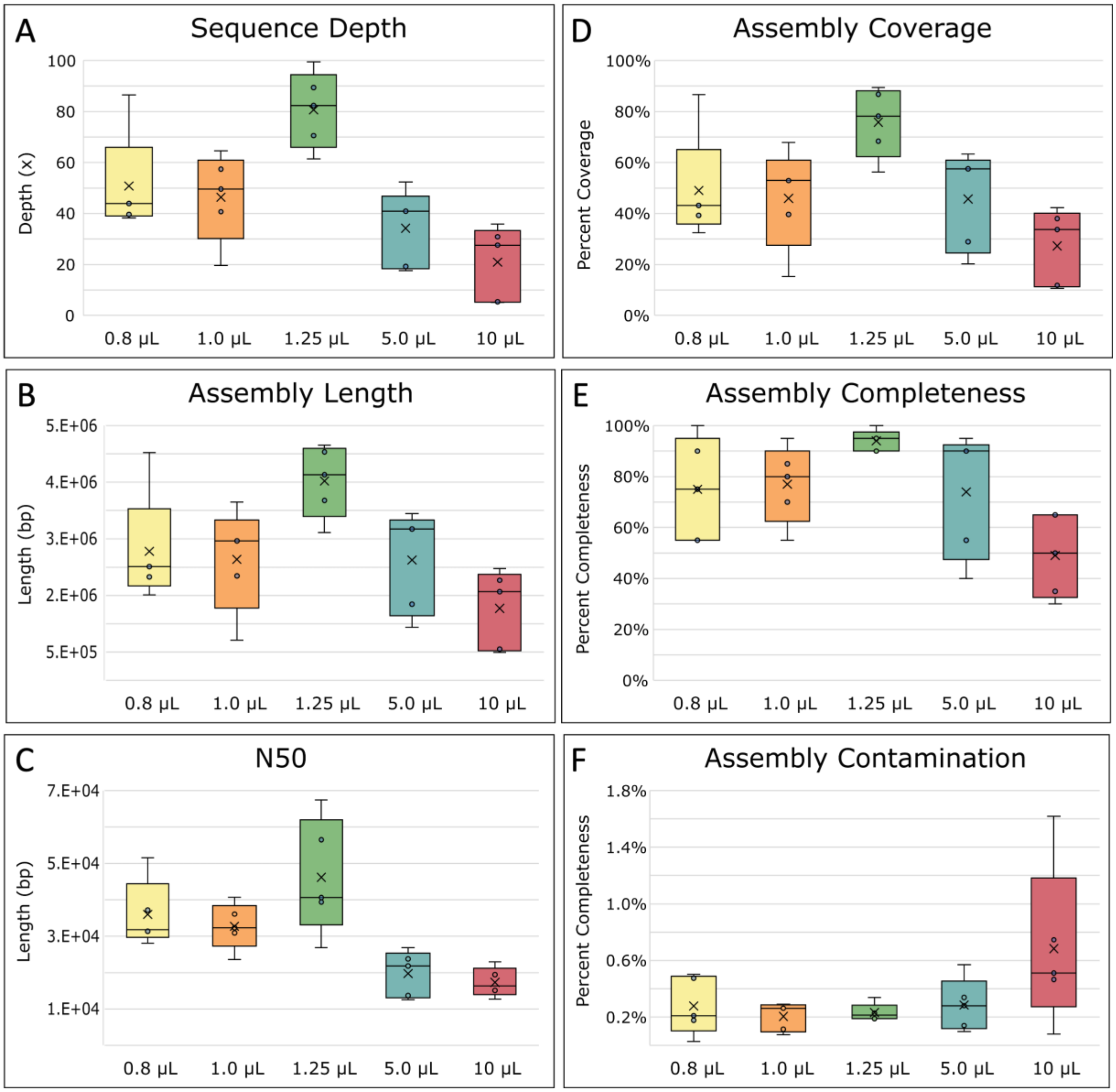


Figure 3.8. Single-amplified (SAG) assembly statistics

A Final sequence depth, calculated as the estimated number of times each base within the genome was sequenced on average. **B** The total average length of the assemblies, **C** N50 average, the minimum contig length needed to support 50% of the genome assembly, and **D** the percent coverage of the assemblies across the *E. coli* MG1655 reference genome, were all determined with QUAST (Mikheenko et al., 2018). **E** The completeness of the assembled genome and **F** the percent of contaminated bases in the assemblies, were determined by MDMCleaner (Vollmers et al., 2022). The boxes middle line represents the median, and the x represents the mean. Five replicates were used for calculation.

3.1.6 MDA volume reduction with DMA

The smallest successful MDA reaction volume in 384-well plates was 0.8 µL, therefore, 0.5 µL MDA reactions were further tested using the DMA (Aquarray, Germany) in order to determine if further volume reduction improved the SAG quality. To combat evaporation, glycerol was added to the MDA master mix at either 5% or 10%. Initial tests in 5 µL found that both concentrations of glycerol still allowed for successful amplification when tested in 5 µL, however, when applied in 1.25 µL reactions, 10% glycerol resulted in much less amplification yield than 5% glycerol (**Figure 3.9**). From MDA reaction kinetics alone, it Is hard to say if this truly has a negative impact on the reaction its self, however, too high of glycerol concentration has been found to prevent the polymerase from accessing the template DNA (Kuznetsova et al., 2014; Ralston, 1990). Therefore, in order to not cause too much molecular crowding, 5% glycerol was used for further tests on the DMA.

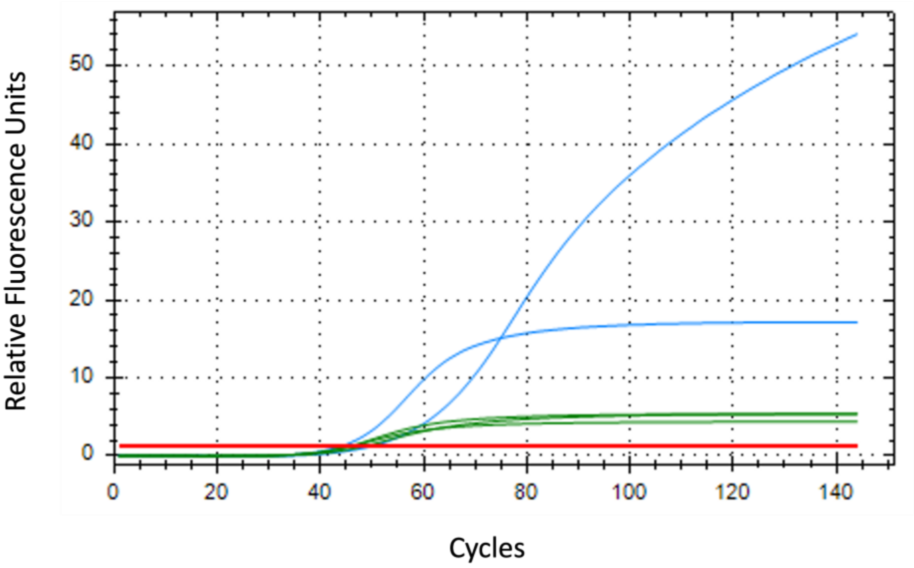


Figure 3.9. MDA reaction kinetics for glycerol testing
MDA master mix with 10% glycerol (green) and 5% glycerol (blue) were tested to determine what effects the glycerol had on the MDA reaction its self. Negative control = red line. Each cycle represents 5 minutes of amplification time. Relative fluorescence units (RFU) refer to the fluorescent signal of SYTO™-13 measured with a real-time thermo-cycler. SYTO™-13 is used to monitor the progression of MDA because it binds to double-stranded DNA as it is amplified.

Initially, *E*. *coli* DNA was tested at 30 and 300 fg total input on the DMA. Due to the small volume size, at the time, three spots were combined for measuring the concentration with Qubit (Thermo Fisher Scientific, USA). As a control, the same DNA concentrations were amplified in a 384-well plate at the same time as the DMA. Both 30 and 300 fg successfully amplified to approximately 0.061 ng and 0.068 ng, respectively. However, approximately 200 nL of the reaction volume was still being lost to evaporation, so further DMA incubations were performed in an incubator with the DMA inside a humidified petri dish (Chakraborty et al., 2022). However, the success of subsequent testing with *E*. *coli* DNA on the DMA was very inconsistent and usually not successful. Still, the ability to amplify genomes from single *E*. *coli* cells on the DMA in 0.5 µL MDA reactions was tested. First, the B.SIGHT™ cell printer (Cytena, Germany) was applied for sorting of single-cells. Bright, fluorescent 5 µM beads were sorted onto the DMA to mimic *E*. *coli* single-cell sorting and prove that cells could first be accurately sorted using the B.SIGHT™ (**Figure 3.10**). Unfortunately, the following tests that applied MDA to actual *E*. *coli* single-cells sorted onto the DMA, were not successful. Recently, the DMA was used to synthesize cDNA from single HeLa cells (Chakraborty et al., 2022) yet, the cDNA only spent approximately one hour on the DMA versus six hours needed for MDA, and amplification was performed off-chip. Therefore, we contribute our failed MDAs on the DMA to evaporation and/or sterical hinderance of the polymerase (Kuznetsova et al., 2014; Ralston, 1990).

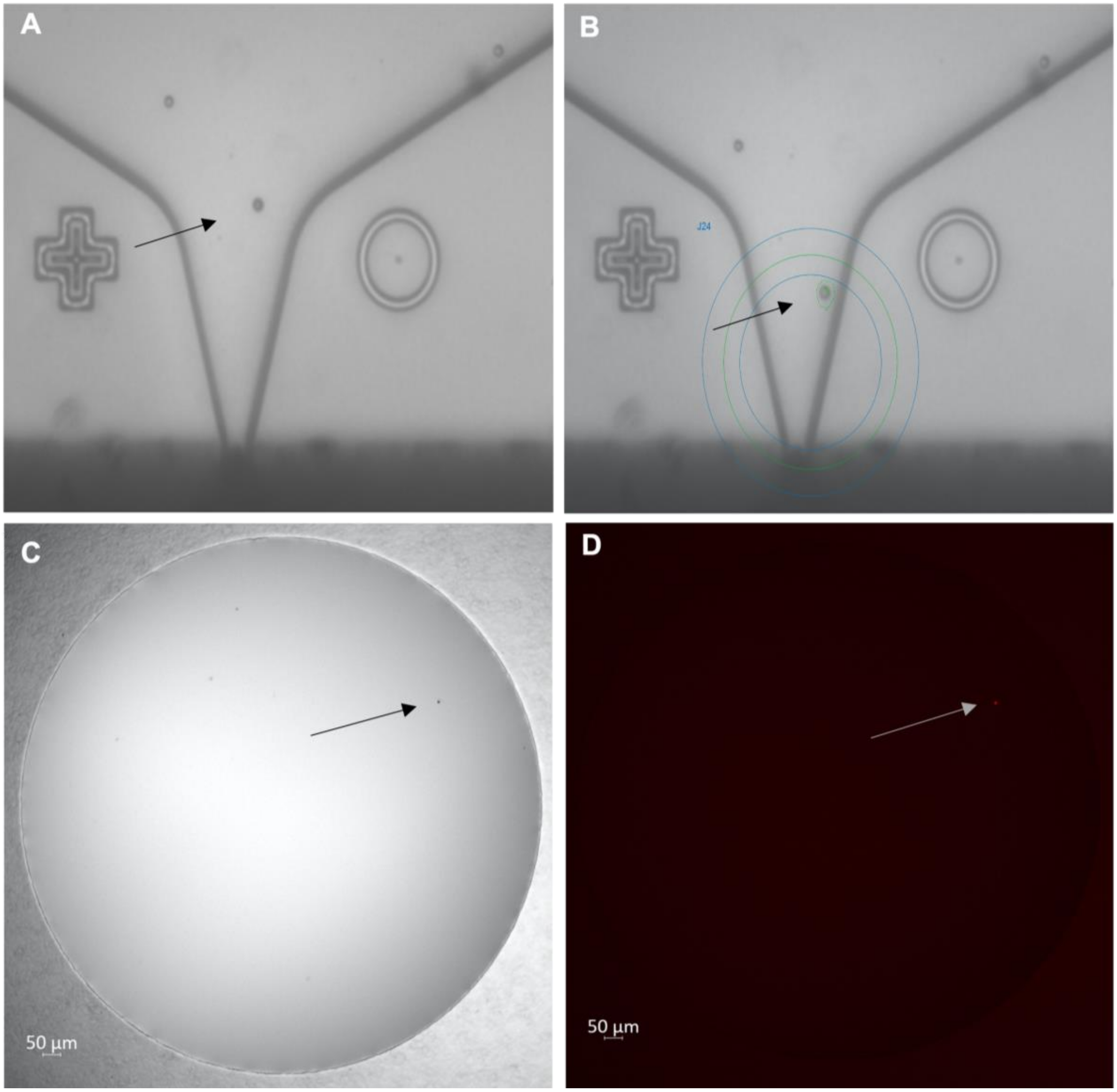


Figure 3.10. Mimicking single-cell sorting onto the DMA

A-B Single 5 µM fluorescent beads were sorted through the single cell printer's 20 µM nozzle dispensing cartridge to mimic single microbial cell isolation onto the DMA. **C** Bright-field and **D** epifluorescence microscopy confirmed a single bead was sorted onto one DMA spot. Microscope images were taken at 50X magnification.

3.1.7 Conclusion and outlook

Overall, these results from this work found several ways to improve upon current challenges in microbial SCG. First, it could be shown that modifying FISH for SCG allowed for enrichment of low abundant, albeit ecologically and biotechnologically important Chloroflexi (Dam et al., 2020). Future targeted-cell sorting studies could benefit by implementing the other improvements made in this dissertation, such as the single-cell printer, diluted lysis buffer, and reduced MDA reaction volumes. Nonetheless, the draft genomes obtained using in solution, fixation-free FISH revealed novel phylogenies, metabolisms, and other physiological characteristics of rare members of the community that would have otherwise been overlooked by conventional metagenomics. It is anticipated that this approach will be used to further reduce the cost of SCG and reveal novel microbial dark matter.

Furthermore, it could be demonstrated that MDA performed in 1.25 µL reaction volumes provides an easy and efficient approach to improve MDA by producing significantly less-biased, less contaminated, and more complete SAGs than standard, larger reaction volumes. Still, further volume reduction could possibly increase genome coverage by ~12-14% (Gole et al., 2013; Nishikawa et al., 2015), however the reproducibility of these picoliter and nanoliter approaches is uncertain since few approaches have been validated outside the original study. This is due to the fact that microfluidic, droplet, and other volume reduction approaches are not as easily accessible or easy to use, and many are not high-throughput enough for hundreds of single cells. In addition, because DNA yield is limited in smaller volumes, some studies have had to perform two rounds of MDA to generate sufficient amounts of products for library preparation (Marcy, Ishoey, et al., 2007; Xu et al., 2016). However, library preparation input requirements have decreased from ug to pg in the last few years (Rinke et al., 2016), so lower DNA yield is no longer much of an issue.

Based on these results, it is a question whether further volume reduction is really necessary. Ultimately, one should gauge for themselves whether the time and costs benefits of volume reduction down to nL and pL reactions makes sense in the scope of their study. Meanwhile, volume reduction in standard 384-well plates and with commercially available cell sorters and liquid dispensers makes this approach more easily accessible to other researchers and already drastically reduces the costs by ~97.5% from the standard 50 µL MDA reaction (**Table 1.1**). It was also found that with this approach, 40X sequence depth is enough for high quality assemblies (**Appendix Figure 1**), compared to the standard >100X depths generally used in microbial SCG (Stepanauskas et al., 2017; Woyke et al., 2011). Further cost reduction could also be achieved by applying this approach to the less expensive WGA-X™ method (**Table 1.1**), seeing that preliminary work in our group finds WGA-X™ to work in 1.25 µL reaction volumes as well. In the end, it is anticipated that the improvements made herein will be of high interest for other single-cell studies and will therefore increase the use of SCG, especially for research focused on elucidating the genomic potential of rare taxa and/or novel microbial dark matter in environmental samples.

3.2 Chapter 2: Improving microbial single-cell transcriptomics

While SCG has been successfully applied to microorganisms, it only provides genomic information, limiting our understanding of microbial function and activity at the single cell level. This, and the fact that the transcriptomic profiles of individual cells vary, even if they are genetically homogenous, underlines the necessity for a robust microbial SCT pipeline. Currently, however, very few methods exist for microbial SCT because of the many physical challenges in working with single microbial cells, thus the goal of this work was to establish a feasible SCT pipeline for prokaryotes.

3.2.1 Improving reverse-transcription for low cell inputs

First, the challenges and limits of working with low inputs of RNA were assessed. Two different reverse transcriptase, M-MLV (Promega, USA) and SuperScript IV (Thermo Fisher Scientific, USA), were used to convert RNA extracted from *E. coli* MG1655 to cDNA. Successful reverse transcription (RT) was examined at 100 ng, 10 ng, 1 ng, and 0.1 ng of total RNA, equivalent to 1 million cells down to 1000 cells. SuperScript IV performed the best since only it allowed for inputs less than 10 ng to be converted to cDNA and amplified with polymerase chain reaction (PCR) (**Figure 3.11**). This could be contributed to the higher temperature threshold that SuperScript IV functions (50°C) compared to MMLV (42°C). Thus, genes with high GC were likely more easily amplified. Total RNA extracted from 10 million to 100 actual *E. coli* cells was then assessed. It was determined that SuperScript IV was limited to 1000 cells, as cDNA from 100 cells could not be successfully amplified. Therefore, the next challenge was to establish a method sensitive enough for single cells.

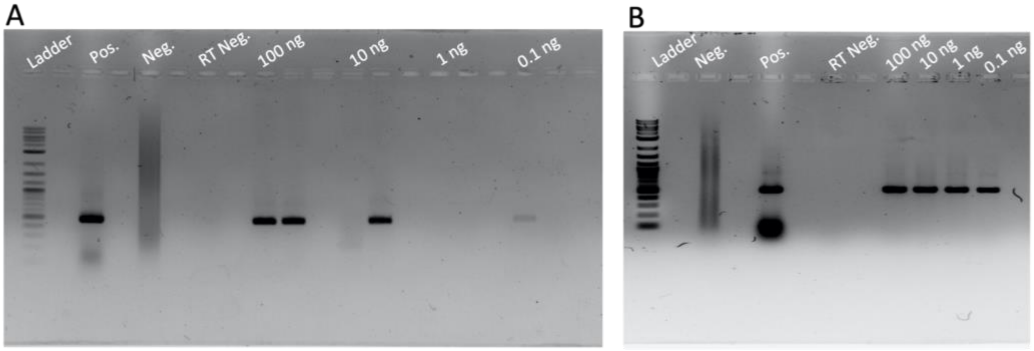


Figure 3.11. Comparison of reverse transcriptase on low input *E. coli* total RNA
Agarose gel electrophoresis of reverse transcriptase PCR results on a 1% agarose gel. **A** M-MLV reverse transcriptase. **B** SuperScript IV reverse transcriptase. Positive control = 20 ng of *E. coli* DNA. Ladder is 1 kb, where the product is ~ 465 bp. The smear in the negative controls is from non-specific binding since 35 cycles were used for PCR.

3.2.2 Oxford Nanopore direct-cDNA

Previously in Chapter 1, the issue of amplification bias in SCG was discussed and assessed, but this too is a problem for transcriptomics in general, not just at the single-cell level. Amplification bias effects the quantification of transcripts and can lead to the detection of false-positive transcripts (Parekh et al., 2016). Currently, Oxford Nanopore Technologies (ONT) has two different long-read RNA-seq protocols, direct-RNA and direct-cDNA, that do not require amplification to reduce amplification bias (Garalde et al., 2018). However, the required RNA input amount for the direct-RNA kit was 500 ng, so it was decided that the direct-cDNA kit, which requires 5x less input, would be assessed and optimized herein for microbial SCT.

First, a polyadenylation protocol (Grünberger et al., 2022) was modified because direct-cDNA requires polyadenylated RNA as input. To increase the sensitivity to lower RNA inputs, polyadenylation was performed in reduced reaction volume, then applied to 4.0 ng and 40 ng of total *E. coli* RNA prior to implementing the direct-cDNA kit. Bioanalyzer results showed that the libraries mostly consisted of rRNA (**Figure 3.12**). Furthermore, sequencing results showed that only 9.9 % and 45.5% of the transcripts could be detected from the 4.0 ng and 40 ng libraries, respectively. Because the 4.0 ng library transcript coverage was so low, it was determined that this method would not be sensitive enough for single-cell levels of microbial RNA. In order for the direct-cDNA and direct-RNA kits to be applicable for microbial SCT in the future, ONT will need to change the reliance on polyadenylated RNA and lower the RNA input required. Thus, further modifications to this approach were outside the scope of this dissertation.

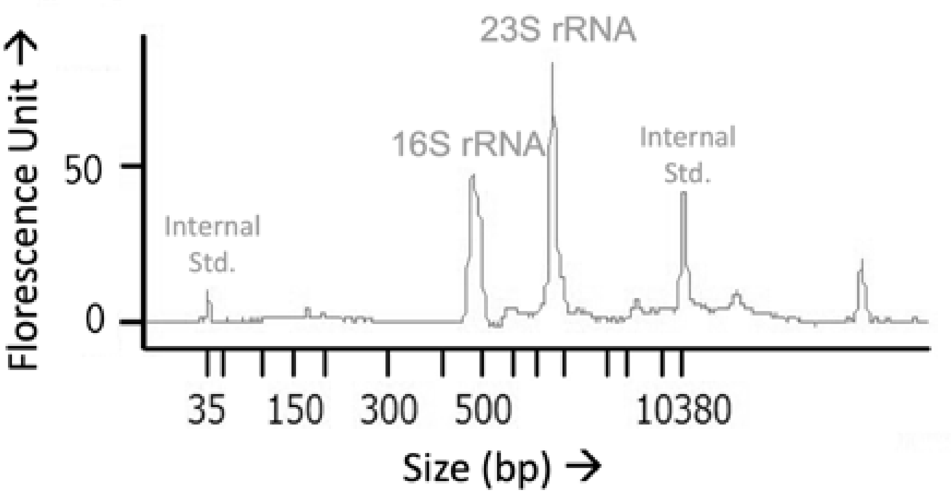


Figure 3.12. ONT direct-cDNA library quality overview
Electropherogram of total RNA library size distribution and quantification assessed with Agilent's Bioanalyzer. The two center peaks represent the 16S and 23S rRNA, which typically make up >95% of a total RNA-seq library.

3.2.3 Optimizing single-cell whole-transcriptome amplification for prokaryotes

Considering that current amplification-free RNA-seq methods are not sensitive enough for single-cell levels of RNA, amplification-based methods are still required but remain lacking for microbial scRNA-seq. The REPLI-g WTA Single Cell kit is an MDA-based, WTA method that uses the same amplification strategy as the REPLI-g Single Cell kit used in Chapter 1, but includes a pre-DNA removal, RT, and ligation step (**Figure 3.13**). Thus, this kit would provide a familiar and accessible approach for microbial SCT. Previously, this MDA-based WTA method was applied to the bacterium *Porphyromonas somerae* (Liu et al., 2019), however, we found their sequence data to be severely contaminated with DNA (discussed in detail below). Therefore, further improvements to this method were pursued in this dissertation work in order to establish MDA-based WTA as a reliable method for microbial scRNA-seq.

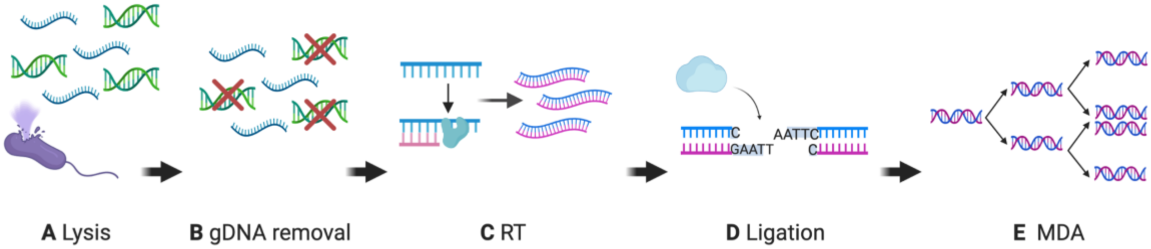


Figure 3.13. REPLI-g Single Cell WTA method overview
A If possible, cells should be immediately isolated and lysed to prevent changes to a cell's transcriptome. **B** DNA is removed after lysis to prevent false-positive transcript counts. **C** Using random primers, RNA is non-specifically primed and reverse transcriptase converts the RNA to cDNA. **D** Due to the nature of the phi29 polymerase used in the following step, cDNA must be ligated to form long templates required for amplification. **E** MDA is employed to exponentially amplify fg of cDNA into ng for subsequent library preparation and sequencing. Made with Biorender.com

Because the lysis buffer provided with the REPLI-g WTA kit does not work on cells with cell walls, like prokaryotes, a different lysis strategy and appropriate lysis buffer were needed to extract the RNA without damaging it. In total, four different lysis methods were assessed. First, the lysis used in the original microbial REPLI-g Single Cell WTA publication (Liu et al., 2019) with and without lysozyme was tested on replicates of single, 10, and 100 *E. coli* cells. But this lysis buffer, with and without lysozyme, did not work for *E. coli* likely because even without lysozyme, the addition of 200 mM KCL and 0.1% Triton X-100 were too harsh for *E. coli* cells and damaged the RNA. Therefore, the lysis conditions were reduced to one freeze-thaw cycle with a lysis buffer that only included 1X PBS + 0.1 U/µL RI + 200 mM DTT. This lysis seemingly allowed for successful WTA of *E. coli* single cells, but a high amount of DNA contamination was found upon analyzing the sequence data, as indicated by reads mapping to intergenic regions along the genome and more uniform coverage (**Figure 3.14)**.

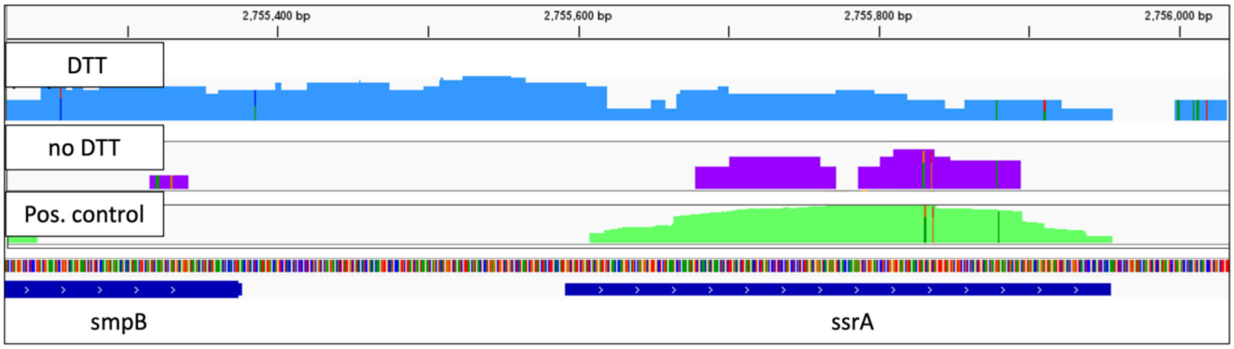


Figure 3.14. Read coverage and density of *smpB* and *ssrA* genes
SCT of cells treated lysed with buffer that contained DTT (blue) and without DTT (purple). Pos. control (green) is bulk extracted RNA treated twice with DNase. Libraries were log scaled and view in Integrative Genomics Viewer.

DNA contamination is a large problem for RNA sequencing in general because it produces false-positives in gene expression data (X. Li et al., 2022). Typically, in bulk RNA-seq, efficient DNA removal can be checked with PCR or qPCR, but this is not possible at the single-cell level due to too little template and RNA degradation. Furthermore, amplification methods for scRNA-seq, like MDA used herein, are very sensitive to contaminating DNA because of the fidelity of the polymerase. Therefore, the raw RNA-seq data from (Liu et al., 2019) was assessed and was found to also be contaminated with DNA (**Appendix Figure 2**). It was found that as little as 0.1 mM of DTT can inactivate the endonuclease DNase I (Hanaki et al., 2000), the enzyme responsible for removing DNA from solutions. When we removed DTT and only 1X PBS + RI was used for lysis, there was no longer detectable DNA contamination in the scRNA-seq data herein (**Figure 3.14**). Thus, it could be concluded that high concentrations of DTT were inhibiting the activity of DNase I previously. These results serve as an important reminder for future studies to check reagent compatibility, even when following published protocols.

Next, additional improvements and modifications were implemented to increase the success of RNA amplification. Microbial RNA has half-life times of 10 min or less (Brennan & Rosenthal, 2021) and with only femtograms of RNA available, the RNA has to be processed or preserved quickly. Therefore, a non-contact liquid dispenser (I.DOT mini, Dispendix) that dispenses reagents into a 384-wellplate in under 5 min, was used to increase the success of WTA while also helping to minimize contamination. Additionally, including RI in the RT master mix is standard in some microbial scRNA-seq methods to reduce RNA degradation (Blattman et al., 2020; Y. Kang et al., 2015, 2011; Kuchina et al., 2021). However, Qiagen keeps their reagent components proprietary and it is thus unknown if RI is already included in the RT master mix, so the addition of RI into the mix was tested here. Because recent methods used Invitrogen's SUPERase - In™ RI (Blattman et al., 2020; Kuchina et al., 2021), the effect of its protection specifically against RNA degradation to Promega's RNasin® was compared. RNasin® RI (Promega) resulted in earlier Cq numbers than SUPERase - In™ RI (Invitrogen), which indicated that SUPERase - In™ likely did not protect RNA from degradation as well as RNasin®. Unknowingly at the time of testing, SUPERase - In™ was previously found to also perform poorly against RNA degradation compared to RNasin® (Probst et al., 2006). Again, this result serves as an important reminder that what works for one method, may not be applicable to others and that reagent compatibility should be tested and compared prior.

Other improvements included increasing the final amplification time to 12 hours and applying the reaction volume reduction approach from Chapter 1 to reduce amplification bias since the SCT method herein also uses MDA. Because RNA-seq data analysis is quantitative, unevenly, over amplified transcripts from amplification bias can cause high false transcript discovery rates and make it difficult to find statistically significant differences between the expression of certain genes (Parekh et al., 2016; Wang & Navin, 2015). As was shown in Chapter 1, reducing reaction volume significantly reduced the over amplification of genome regions, thus here the scWTA reactions were lowered 5-fold, from the standard 60 µL down to 12 µL, to achieve the same effect. This also helped to reduce the costs of the WTA from approximately 55 USD to 11 USD per sample. However, it could not be said definitively if reducing the WTA reaction volume improved the bias because direct comparisons could not be made since the 60 µL reactions did not work. Though future work should assess the bias between other, smaller volumes to see if the effect is similar to that in Chapter 1. Further reduction was also tested in 3 µL WTA reaction volumes, but the number of successfully amplified cells was less, possibly due to the evaporation of the low volumes (≤ 1 µL) of REPLI-g WTA's initial steps (**Table 2.2**). Further volume reduction may be accomplished in the future by optimization of the reagent concentrations and timing between reagent dispensing.

3.2.4 Modified single-cell WTA generates reliable scRNA-seq data

The modified REPLI-g WTA method was then applied to *E. coli* cells that were subjected to heat-shock at 50°C for 10 min and non-treated cells to validate its ability to differentiate between differently treated cells. The number of heat-shock scWTA samples sequenced included 21 single cells, 4-10 cells, and 4-100 cells. For non-treated samples, 23 single cells, 6-10 cells, and 6-100 cells were sequenced. Two bulk RNA-seq samples that were either amplified (positive control) or not amplified with REPLI-g WTA (bulk) were also sequenced from both heat-shock and non-treated cells (1.2 million cells each) as controls. Checks for DNA contamination were conducted for all sequenced samples cells and can be found in **Appendix Figure 3** for heat-shock samples and **Appendix Figure 4** for non-treated samples. Principal component analysis (PCA) of all samples, based on the top 300 variant genes, clustered samples based on cell number (**Appendix Figure 5A**), therefore, cell number was included into the DESeq2 design along with condition for analysis between all samples.

Overall, the number of transcripts detected increased as the cell number increased (**Figure 3.15A**). The average gene coverage was 7.2% for single-cells, 10.6% for 10-cells, 35.4% for 100-cells, 82.4% for bulk, and 17.4% for positive controls (**Appendix Table 6**). Furthermore, approximately 50 more genes were detected within heat-shocked cells than non-treated on average (**Figure 3.15A**). A similar result was also reported for single *E. coli* cells subjected to heat-shock with a different scRNA-seq method (Kuchina et al., 2021), and is likely a result of overexpression of regulatory genes responding to stress (Gunasekera et al., 2008).

Even though our gene detection from single-cells may not be enough to fully represent a microbial transcriptome (Haas et al., 2012), the detection efficiency is realistic based on recently published results of *E. coli* scRNA-seq (Blattman et al., 2020; Kuchina et al., 2021). Other previously established methods reported detecting transcripts covering up to 75% (Liu et al., 2019) and 99% (Wang et al., 2015) of the genome from a single *E. coli* cell, but this is questionable when considering that most microbial transcripts are estimated to only be present in less than one copy per cell (Imdahl & Saliba, 2020). It is more likely that these relatively high gene capture rates were due to false-positives caused by DNA contamination, which is supported by the fact that we could confirm DNA contamination in the data from (Liu et al., 2019) (**Appendix Figure 2**).

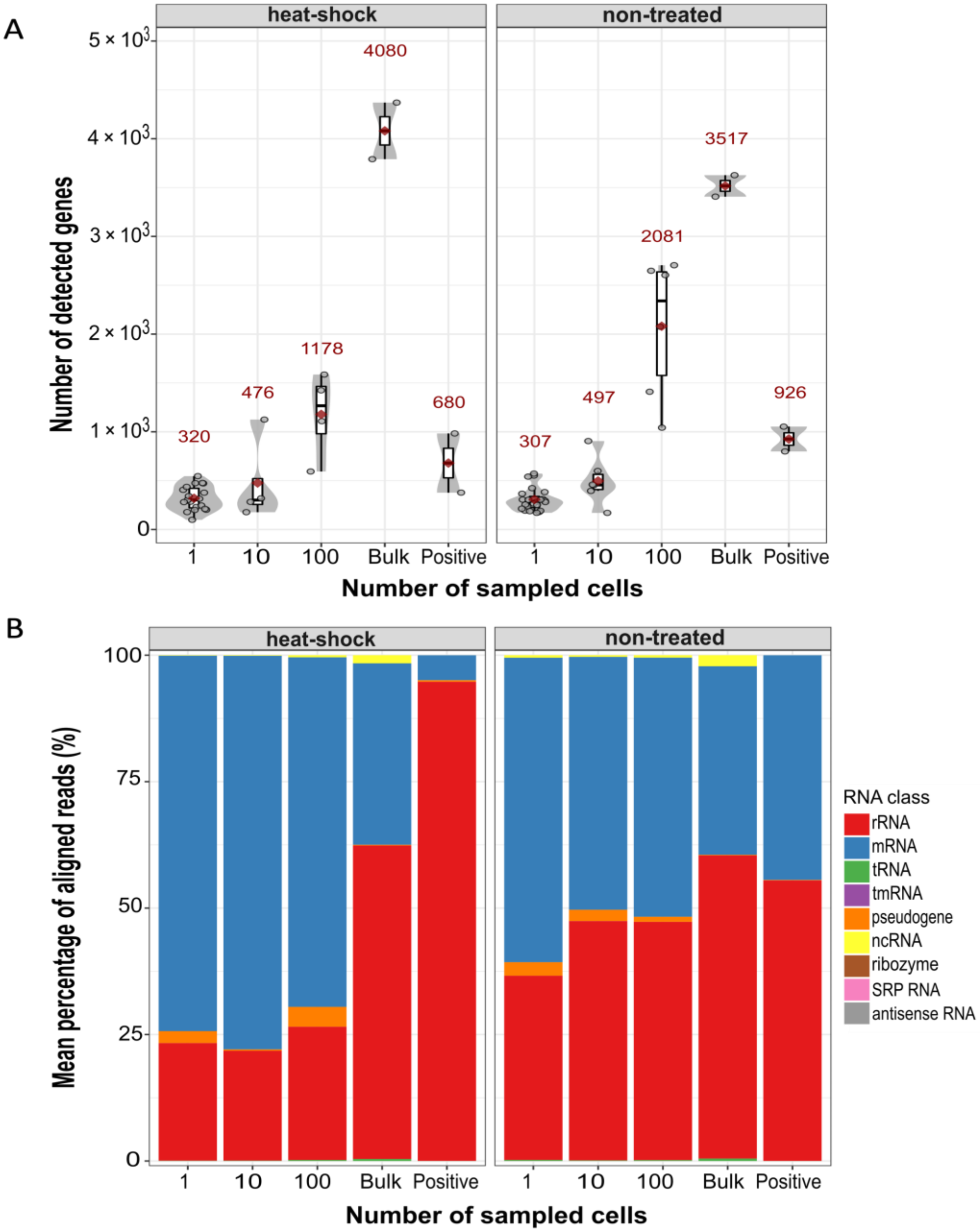


Figure 3.15. Transcript characteristic summary for heat-shock and non-treated samples
A Average number of transcripts, detected if the gene has >5 transcripts mapped per a sample. **B** Relative proportions of transcripts for each RNA class type.

Surprisingly, the number of rRNA transcripts detected in the single, 10, and 100 cells was much lower (**Figure 3.15B**) than what would be expected (~95-99%). But because the positive MDA-amplified control sample had >90% rRNA, this effect was possibly contributed to insufficient denaturization of the rRNA secondary and tertiary structures. If this is the case, then bias against other non-linear genes would exist, but significant differences in counts from genes with known secondary structures such as *ompF*, *deaD*, and *purA* (Del Campo et al., 2015) were not found. Next, genes that were missing in both the bulk RNA-seq and scRNA-seq method were compared (**Figure 3.16**). As seen in **Figure 3.16A**, both bulk RNA-seq and MDA-based scRNA-seq are biased against small transcripts, which a known issue with RNA-seq in general (Oshlack & Wakefield, 2009). The more severe bias in the scRNA-seq samples may be contributed to the phi29 polymerase since it requires long templates for synthesis (Gadkar & Filion, 2012). Based on the GC% of the missing genes, there was no significant bias associated to GC content between all genes and genes missed with MDA scRNA-seq (**Figure 3.16B**). But there was a significant bias against lower GC% genes in bulk RNA-seq. However, considering that the rRNA operon in *E*. *coli* is >4500 bp long and that the GC% is average (~54%), these results still do not explain why lower rRNA percentages were detected. Therefore, the bias against rRNA in the scRNA-seq data at the present time is contributed to differences in cell isolation and lysis since the MDA positive control samples had high percentages of rRNA (**Figure 3.15**). Future work should determine if these differences are truly contributed to single-cell sample processing step by improving upon the cell lysis and denaturation step.

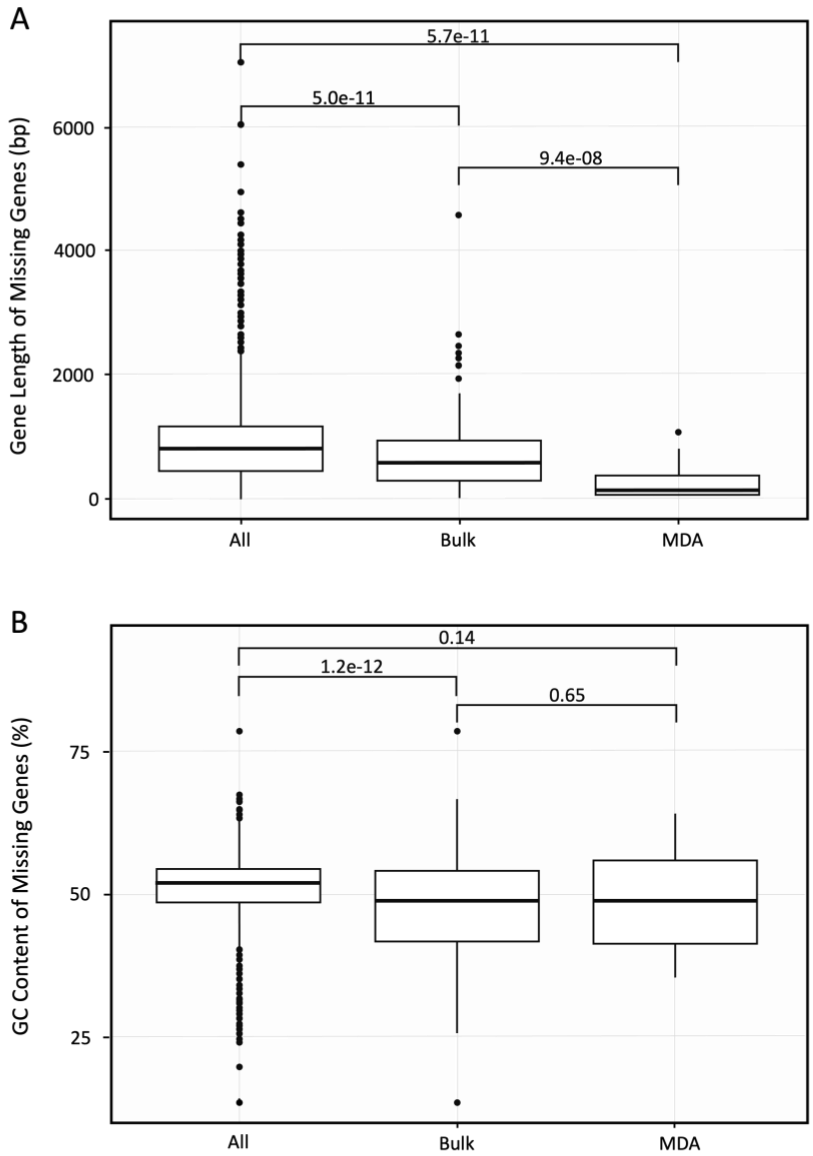


Figure 3.16. Gene length and GC content bias in RNA-seq
A Gene length bias of missing genes. **B** GC % bias of missing genes. Significant differences are indicated by the top brackets (alpha=0.05). All represents the average length for all genes in the *E*. *coli* genome.

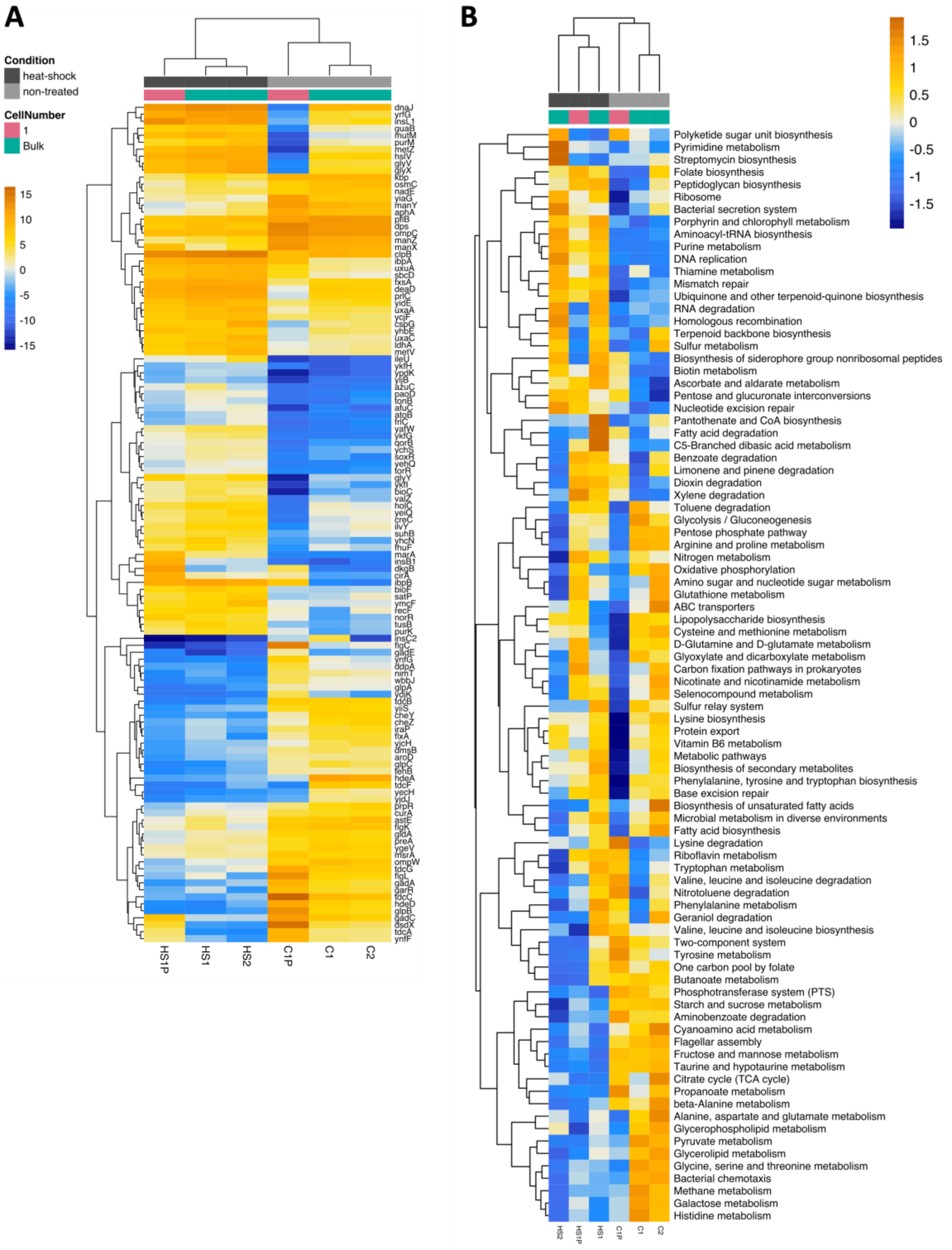


Figure 3.17. Pseudo-bulk versus bulk RNA-seq analysis

A Heatmap of the 120 most significantly expressed genes. **B** Enriched KEGG pathway analysis. Normalized data was scaled with z-scores. Samples are as follows: HS1P; heat-shock pseudo-bulk, HS1; heat-shock bulk 1, HS2; heat-shock bulk 2, C1P; control pseudo-bulk, C1; control 1, C2; control 2.

All single-cells gene counts were then combined to create a pseudo-bulk sample for each treatment type to be compared to their respective true bulk sample as a way to confirm the reliability of the MDA-based scWTA method used herein (**Figure 3.17**). Gene coverage increased to 99.76% for the heat-shock pseudo-bulk sample and 94.34% for the non-treated. It was confirmed, based on the 120 most significantly expressed genes (**Figure 3.17A**) and with KEGG pathways enrichment analysis (**Figure 3.17B**), that pseudo-bulk samples cluster with their respective “true” bulk RNA-seq sample by treatment type.

3.2.5 Heterogeneous single-cells still differentiate between treatment type

Unbiased clustering of only the single-cells, based on the 300 genes with the most variance, did not delineate cells between treatment (**Appendix Figure 5B**). Furthermore, heat-shock cells G17Sep and F17April, as well as non-treated cells F09Sep, E10Sep, E08Sep, and F10Sep were identified as outliers based on the PCA plot and thus, removed from further analysis. A closer look into the transcription of the eight most upregulated heat-shock genes in heat-shock treated single cells found several cells upregulating many of the heat-shock genes at once (**Figure 3.18**). However, some non-treated cells were also found to transcribe these heat-shock genes, namely *hslO*, *plsB*, *dnaE*, and *hflX*. But, between the two treatment types, *hflX*, *dnaE*, *fkpA*, *Lon*, and *rpoD* were considered significantly differentially upregulated (padj < 0.05; log2FoldChange > 0) (**Appendix Table 6**). Still, the expression of heat-shock genes in non-treated cells warranted further investigation. The majority of heat-shock genes are molecular chaperones that help repair misfolded and/or damaged polypeptides during cellular stress (Feder & Hofmann, 1999). Reports on the expression of heat-shock genes in *E. coli* find that many of these genes are also upregulated under oxidative and starvation stress (Díaz-Acosta et al., 2006; Ngan et al., 2021; S. Wang et al., 2009; Winter et al., 2005), which cells typically experience at high cell densities (i.e. late exponential phase) (Yoon et al., 2003). It is then possible that our non-treated cells were experiencing stress related to population overgrowth. Therefore, differences in the transcription of some heat-shock genes between treated and non-treated cells at the single-cell level could not be well defined since heat-shock genes seem to have a more general role in stress response.

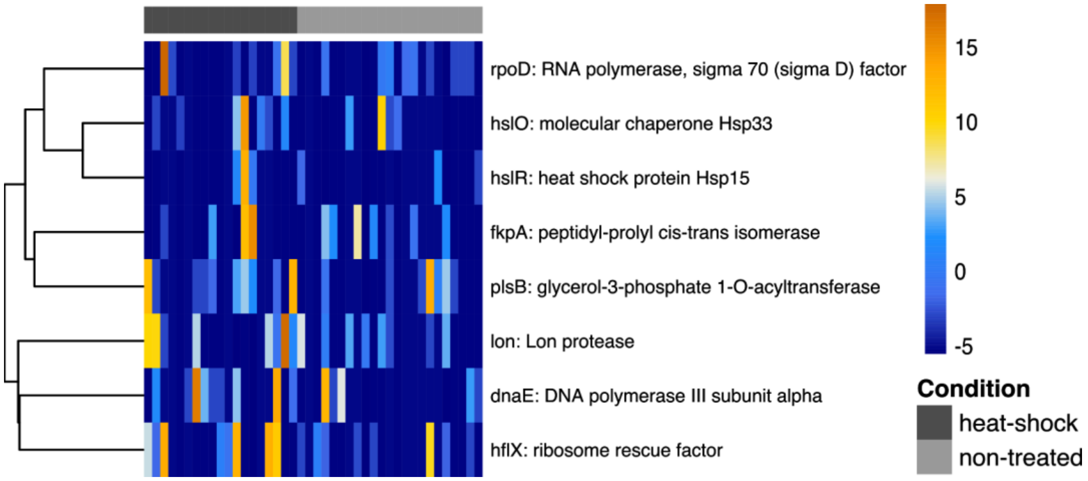


Figure 3.18. Highest transcribed heat-shock genes from single-cells
Normalized-transformed read counts from heat-shock treated single cells were ordered from highest to lowest and the top eight heat-shock genes selected for comparison with non-treated cells. Counts were scaled with $\log_2$.

In order to further understand what were the factors controlling separation of samples by treatment type in **Figure 3.17**, KEGG pathway enrichment heat maps were analyzed in more detail. Based on **Figure 3.17B**, heat-shock samples upregulated DNA repair pathways such as, DNA replication (eco03030), Mismatch repair (eco03430), Nucleotide exclusion Repair (eco03420), as well as Ubiquinone and other terpenoid-quinone biosynthesis (eco00130), and purine (eco00230) and thiamine metabolism (eco00730) pathways, when compared to non-treated cells. This could also explain why specifically looking at heat-shock genes was not as informative as most of the single-cells cells were already under going repair after the initial heat-shock response. On the other hand, non-treated single-cells upregulated standard metabolic pathways for growing cells, such as Phosphotransferase system (eco02060), Starch and sucrose metabolism (eco00500), Aminobenzoate degradation (eco00627), and Fructose and mannose metabolism (eco00051). Additionally, non-treated single-cells differentially upregulated the Flagellar assembly pathway (eco02040) which is expected for fast growing *E. coli* cells (Sim et al., 2017).

3.2.6 Microbial SCT reveals rare and unique activity

Next, enriched pathways of the single-cells were compared to their respective bulk RNA-seq samples, again from **Figure 3.17B**, to understand what genes/pathways were missed with bulk RNA-seq methods. In order for a pathway to be considered present in the single cells but below detection the bulk samples, the pathways z-score had to be higher than 0 for the single-cell sample and 0 or less for the bulk. In heat-shocked single cells, the pathways involving D-Glutamine and D-glutamate metabolism (eco00471), Glyoxylate and dicarboxylate metabolism (eco00630), Carbon fixation pathways in prokaryotes (eco00720), Oxidative phosphorylation (eco00190), Arginine and proline metabolism (eco00330), Glutathione metabolism (eco00480), and ABC transporters (eco02010), were upregulated compared to their respective bulk controls. Interestingly, the non-treated single cells upregulated several individual pathways that are part of the large biosynthesis of secondary metabolites KEGG pathway (eco01110) when compared to the non-treated bulk control samples. These included Polyketide sugar unit biosynthesis (eco00523), Biosynthesis of siderophore group non-ribosomal peptides (eco01053), Biotin metabolism (eco00780), Ascorbate and aldarate metabolism (eco00053), Pentose and glucuronate interconversions (eco00040), Lysine degradation (eco00310), Riboflavin metabolism (eco00740), and Dioxin degradation (eco00621). This indicated that some cells were likely producing secondary metabolites. By taking a look at the normalized-transformed count data, we found evidence that a small subset of single-cells transcribed genes required for the synthesis of the secondary metabolite, enterobactin, which was lowly expressed in comparison to the bulk control samples (**Figure 3.19A**). Additional comparison of fold change differences between pseudo-bulk and true bulk samples of significantly expressed enterobactin genes (padj <0.05) further confirmed that genes *entB*, *entD*, and *fepA* had higher expression (log2FoldChange < 0) in pseudo-bulk samples overall (**Appendix Table 7**). In a clinical setting, pathogenic *E. coli* produce enterobactin, a siderophore, to decrease iron availability needed for the host's immune response and provide iron needed for pathogen growth (Golonka et al., 2019). In isogenic populations, however, siderophore production is still important in iron-deficient conditions, but also for signaling biofilm formation (May & Okabe, 2011) and responding to oxidative stress response

(Peralta et al., 2016). Again, this finding further provides evidence that some non-treated cells were likely experiencing the effects of high cell density as a result of a fastly growing population.

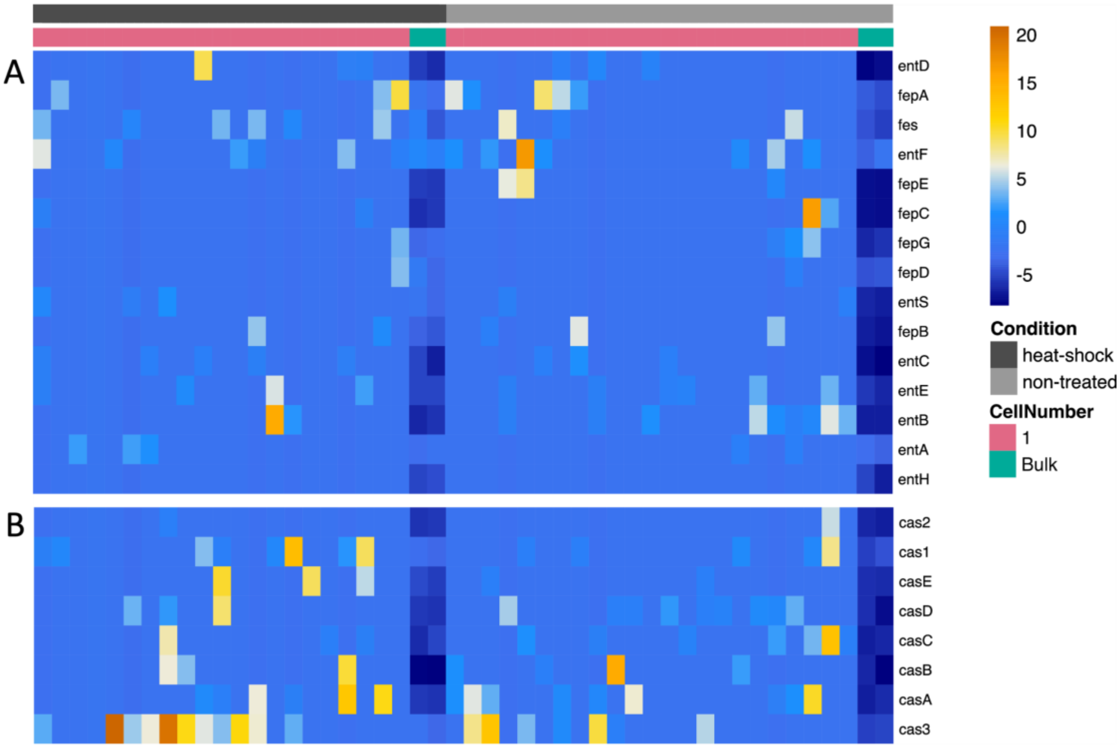


Figure 3.19. Bulk vs. scRNA-seq transcription of Enterobactin and Type I CRISPR Cas genes
A Enterobactin synthesis genes. **B** Type I CRISPR system genes. Normalized counts were scaled with $\log_2$.

But why would only a small sub-population of cells express enterobactin genes? One possible explanation for heterogenous expression of enterobactin could be due to division of labor and/or bet-hedging tactics that microorganisms use to increase the populations fitness (Morawska et al., 2022). The metabolic costs of producing secondary metabolites, like siderophores, are costly for the cell (Lv et al., 2014), thus a population could benefit by dividing tasks that benefit the entire populations growth amongst specialized individual cells and allow them to be ready to react to environmental changes (Ackermann, 2015). A recently deposited preprint found evidence using microscopy that at the single-cell level, siderophore production in *Pseudomonas aeruginosa* is heterogeneously expressed when intracellular iron stocks are greater in most cells and homogenously expressed over time when iron becomes limiting for most cells (Mridha & Kümmerli, 2021). Furthermore, the authors also found that at low iron levels, some level of siderophore expression was always on. These findings support two ideas about phenotypic heterogeneity of enterobactin herein. One, that some specialized cells in isogenic *E. coli* populations likely maintain enterobactin production to react to changes in environmental iron concentrations. And two, that heterogeneity observed between cells is also due to differences in metabolic states of individual cells. In the future, applying scRNA-seq to temporal studies will help provide further insight into the regulation of other secondary metabolites and signaling mechanisms that may be missed with bulk RNA-seq.

Additionally, several single-cells transcribing many of the type I CRISPR-Cas system genes were found, which was more emphasized in heat-shock single cells (**Figure 3.19B**). Specifically, all *cas* genes except for *cas2* were considered significantly upregulated (padj < 0.05; log2FoldChange <0) for pseudo-bulk compared to true bulk samples (**Appendix Table 7**). Typically, type I CRISPR-Cas is largely known for its immune response which destroys invading plasmids and viruses (Barrangou & Marraffini, 2014; Sorek et al., 2008), however, there is a growing body of evidence that suspects that these systems are not only for defense, but also for endogenous gene regulation (Bozic et al., 2019). This implied that type I CRISPR-Cas regulation may be important for stress-response experienced by the single-cells herein. In fact, the regulation of c*as3* was previously found to be connected to the presence and transcription of heat-shock gene *htpG* (Yosef et al., 2011). However, because these systems are not typically expressed under standard growth conditions, it has been difficult to identify and understand their "non-canonical" mechanisms (Bozic et al., 2019), especially at the single-cell level. The fact that we found several single cells expressing *cas* genes at much higher levels, supports that previous bulk analysis has largely been unable to detect these genes. Further support for this idea recently confirmed *via* single-cell time lapse microscopy, that only small subpopulations of *E. coli* cells express CRISPR-Cas systems as a quick response to threats (McKenzie et al., 2022), suggesting why this response may not always be found in bulk RNA-seq data. This is the first-time that upregulation of Type I CRISPR Cas genes have been reported in microbial scRNA-seq data, highlighting how important scRNA-seq is for identifying rare and poorly characterized cell responses.

3.2.7 Conclusion and outlook

The dissertation work overcame several difficulties associated with obtaining reliable microbial scRNA-seq data by modifying an existing MDA-based eukaryotic scWTA method. Challenges were circumvented by implementing a lysis procedure specifically for microbial cells, preventing DNA contamination, reducing reagent dispensing times and reaction volumes, as well as including RNase inhibitors to prevent RNA degradation. By comparing heat-shocked single-cell expression results to non-treated control cells, the heterogenous nature of single cells could be confirmed, while cellular states related to heat-shock response and actively growing non-treated cells could be validated separately. Importantly, pseudo-bulk RNAseq data generated from the single-cells confirmed that MDA-based scRNA-seq accurately maintains the global gene transcription profiles seen in the bulk RNA-seq data for both heat-shock and non-treated cells. Furthermore, with this method, rare and unexpected cellular states were uncovered, highlighting that MDA-based microbial scRNA-seq has a clear advantage compared to bulk RNA-seq data when it comes to detecting low abundant, albeit important transcripts in isogenic microbial cultures. However, it should be emphasized that microbial single-cell RNA-seq is not meant to replace bulk RNA-seq as they both have a purpose by providing two different types of information. Here we show that combining the two approaches helped to tease apart functions that had stayed hidden in the bulk transcriptomic data. Because single-cell expression data is noisy and struggles with transcript capture efficiency, a control bulk RNA-seq sample is always essential to ensure that the scRNA-seq data is statistically sound (Imdahl et al., 2020; Squair et al., 2021).

It should be noted that currently, there are some limitations to the MDA-based method used herein in comparison to recently published combinatorial indexing approaches (Blattman et al., 2020; Kuchina et al., 2021) (**Table 1.2**). These methods are capable of sequencing >20,000 cells at once making them more high-throughput. Because of this, they are currently better at identifying sub-populations of cells with different functions. Second, costs per cell (not considering sequencing costs) is ~3000-10,000X cheaper, since library preparations don’t have to be conducted individually for each cell. It is anticipated that replacing the random hexamer reverse-transcriptase primers with barcoded primers will allow all samples to be used for a single library preparation, driving costs down at least 10X and increasing throughput for the MDA-based scRNA-seq method. Additionally, further improvements can also be made, for example removing rRNA to reduce the need for high sequence depth. Since the RNA was very sensitive to timing, methods that can be applied after cDNA amplification (Gu et al., 2016; Prezza et al., 2020) would likely work best. And, as mentioned previously, further volume reduction may be possible with optimization of the reagent concentrations and dispensing time in order to reduce costs.

Besides throughput and cost factors, there are also some benefits to using MDA-based scRNA-seq compared to the combinatorial approaches that treat all cells as one sample (Blattman et al., 2020; Kuchina et al., 2021). Because cells herein are sorted individually, like in MATQ-seq (Imdahl et al., 2020), less cells are lost during sample processing. Kuchina et al. (2021) reported that only ~25% of cells could be retained throughout the workflow (Kuchina et al., 2021). The authors were still able to identify rare subpopulations (~0.01%), but the effect of cell loss will be more pronounced in a more diverse microbial community. Single-cell isolation also reduces cross-contamination comparatively. Blattman et al. (2020) estimated that up to 5% of transcripts within a single-cell could be derived from other cells when performing species-mixing experiments (Blattman et al., 2020). Additionally, the hands-on time for the combinatorial approaches is longer since the methods require several washing, filtration, and pipetting steps. With the use of the liquid dispenser herein, each reagent dispensing step could be reduced to <5 min. Lastly, soon, microbial single-cell research will move towards capturing multi 'omics data from single-cells, such as simultaneous genome and transcriptome analysis (Song et al., 2019). This approach will be crucial if single-cell RNA-seq is to be applied in the future to environmental samples, since many of the microorganisms currently do not have a reference genome, which is needed for accurate transcript counting. The combinatorial scRNA-seq approaches will likely not be applicable to multi-omics studies because the scWTA reaction occurs within the cells themselves and therefore the cells are no longer useable for other analyses. It is anticipated, however, that MDA-based multi 'omic analysis will soon be possible for microorganisms considering that this approach is already available for eukaryotes (Korfhage et al., 2015).

In summary, the microbial MDA-based scRNA-seq method herein provides valuable information regarding the heterogeneity of microbial single-cells and highlights the promising future this method has for microbial single-cell research. With future improvements to throughput and costs, this method is anticipated to not only be valuable to cultured organisms, but also in elucidating the function of MDM from diverse environmental samples.

4 Future Work

SCG and SCT together have the potential to bring more clarity to the nature of MDM and their metabolic potentials and will enable us to provide information on the structure and dynamics of natural microbial populations in all kinds of environments. Nevertheless, there still remain many areas for further improvement and advancement that can build off the work of this dissertation.

For instance, researchers are continuously publishing new algorithms and computational solutions to overcome issues with amplification bias, drop-outs, and false-positive transcripts, but these methods have all been bench marked on eukaryotic single cells (Adil et al., 2021; Wolfien et al., 2021). It can be expected that as more and more microbial scRNA-seq studies are published, new data solutions based specifically off microbial studies will advance this field further. Another area of research that has yet to be explored in microbial single-cell analysis is the field of single-cell multi ‘omic analysis. This field has taken off in the last few years for eukaryotes now that scRNA-seq has been around for some time (Chen et al., 2021; Macaulay et al., 2017; Song et al., 2019). Soon, this analysis should become possible for microorganisms and provide useful for future studies wishing to assign function to novel MDM. Lastly, the ultimate goal is to forego any the need for amplification to avoid issues that arise from amplification bias. Currently, one solution is to use unique molecular identifiers (UMI) early on the workflow, so that in the end all samples can be pooled into one sample with the hope that there is enough material for library preparation (Chen et al., 2018; Parekh et al., 2016). However, this method is not truly bias-free because it still depends on the successful ligation of the UMI’s to the nucleic acids. Thus, future technologically advancements will hopefully reduce the amount of material needed for library preparation and sequencing down from pg to fg.

Overall, the results of this work are anticipated to allow for more widespread use of the improved single cell ‘omics methods by reducing cost and using methods that are easily accessible to other single-cell research groups. Through this work, further understanding of MDM diversity and function in the environment can be achieved.

5 References

Abou Seeda, M. A., Yassen, A. A., & Abou El-Nour, E. Z. A. A. (2017). Microorganism as a tool of bioremediation technology for cleaning waste and industrial water. *Bioscience Research*, *14*(3), 633–644.

Ackermann, M. (2015). A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews. Microbiology*, *13*(8), 497–508.

Acuña-Amador, L., Primot, A., Cadieu, E., Roulet, A., & Barloy-Hubler, F. (2018). Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of Porphyromonas gingivalis reference strains. *BMC Genomics*, *19*(1), 54.

Adil, A., Kumar, V., Jan, A. T., & Asger, M. (2021). Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. *Frontiers in Neuroscience*, *15*, 398.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, *31*(6), 533–538.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Loman, N. J., Andersson, A. F., & Quince, C. (2013). *CONCOCT: Clustering cONtigs on COverage and ComposiTion*. http://arxiv.org/abs/1312.4038

Amann, R. I., Krumholz, L., & Stahl, D. A. (1990). Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *Journal of Bacteriology*, *172*(2), 762–770.

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., Thomas, B. C., Singh, A., Wilkins, M. J., Karaoz, U., Brodie, E. L., Williams, K. H., Hubbard, S. S., & Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*, 1–11.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., … Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, *9*(1), 75.

Bansal, V. (2017). A computational method for estimating the PCR duplication rate in DNA and

RNA-seq experiments. *BMC Bioinformatics*, *18*(Suppl 3), 43.

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, 201711842.

Barrangou, R., & Marraffini, L. A. (2014). CRISPR-cas systems: Prokaryotes upgrade to adaptive immunity. In *Molecular Cell* (Vol. 54, Issue 2, pp. 234–244). Mol Cell.

Batani, G., Bayer, K., Böge, J., Hentschel, U., & Thomas, T. (2019). Fluorescence in situ hybridization (FISH) and cell sorting of living bacteria. *Scientific Reports*, *9*(1), 1–13.

Bäumer, C., Fisch, E., Wedler, H., Reinecke, F., & Korfhage, C. (2018). Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification. *Scientific Reports |*, *8*, 7476.

Bayer, K., Jahn, M. T., Slaby, B. M., Moitinho-Silva, L., & Hentschel, U. (2018). Marine sponges as Chloroflexi hot-spots: Genomic insights and high resolution visualization of an abundant and diverse symbiotic clade. *BioRxiv*, 328013.

Beale, D. J., Karpe, A. V., & Ahmed, W. (2016). Beyond metabolomics: A review of multi-omics-based approaches. *Microbial Metabolomics: Applications in Clinical, Environmental, and Industrial Microbiology*, 289–312.

Becraft, E. D., Dodsworth, J. A., Murugapiran, S. K., Ohlsson, J. I., Briggs, B. R., Kanbar, J., De Vlaminck, I., Quake, S. R., Dong, H., Hedlund, B. P., & Swingley, W. D. (2016). Single-Cell-Genomics-Facilitated Read Binning of Candidate Phylum EM19 Genomes from Geothermal Spring Metagenomes. *Am Soc Microbiol*.

Bellais, S., Nehlich, M., Ania, M., Duquenoy, A., Mazier, W., van den Engh, G., Baijer, J., Treichel, N. S., Clavel, T., Belotserkovsky, I., & Thomas, V. (2022). Species-targeted sorting and cultivation of commensal bacteria from the gut microbiome using flow cytometry under anaerobic conditions. *Microbiome*, *10*(1), 1–17.

Belotserkovskii, B. P., Johnston, B. H., Gaillard, C., & Strauss, F. (1996). Polypropylene tube surfaces may induce denaturation and multimerization of DNA. *Science*, *271*(5246), 222–223.

Björnsson, L., Hugenholtz, P., Tyson, G. W., & Blackall, L. L. (2002). Filamentous Chloroflexi (green non-sulfur bacteria) are abundant in wastewater treatment processes with biological nutrient removal. *Microbiology*, *148*(8), 2309–2318.

Blainey, P. C. (2013). The future is now: Single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews*, *37*(3), 407–427.

Blainey, P. C., Mosier, A. C., Potanina, A., Francis, C. A., & Quake, S. R. (2011). Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PloS One*, *6*(2). https://doi.org/10.1371/journal.pone.0016626

Blattman, S. B., Jiang, W., Oikonomou, P., & Tavazoie, S. (2020). Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nature Microbiology*, *5*(10), 1192–1201.

Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., Suarez Duran, H. G., de los Santos, E. L. C., Kim, H. U., Nave, M., Dickschat, J. S., Mitchell, D. A., Shelest, E., Breitling, R., Takano, E., Lee, S. Y., Weber, T., & Medema, M. H. (2017). antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research*, *45*(W1), W36–W41.

Bossert, M., Kracht, D., Scherer, S., Landstorfer, R., & Neuhaus, K. (2018). Improving the Reliability of RNA-seq: Approaching Single-Cell Transcriptomics To Explore Individuality in Bacteria. In *Lecture Notes in Bioengineering* (pp. 181–198). Springer, Cham.

Bozic, B., Repac, J., & Djordjevic, M. (2019). Endogenous gene regulation as a predicted main function of type I-E CRISPR/Cas system in *E. Coli*. *Molecules* , *24*(4).

Brennan, M. A., & Rosenthal, A. Z. (2021). Single-Cell RNA Sequencing Elucidates the Structure and Organization of Microbial Communities. *Frontiers in Microbiology*, *12*, 1942.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208–211.

Bushnell, B. (2014). *BBtools software package* (36.84).

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., & Gordon, J. I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Chakraborty, S., Luchena, C., Elton, J. J., Schilling, M. P., Reischl, M., Roux, M., Levkin, P. A., & Popova, A. A. (2022). "Cells-to-cDNA on Chip": Phenotypic Assessment and Gene

Expression Analysis from Live Cells in Nanoliter Volumes Using Droplet Microarrays. *Advanced Healthcare Materials*, *11*(12), 2102493.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* .

Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., & Chen, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biology*, *19*(1).

Chen, Y., Song, J., Ruan, Q., Zeng, X., Wu, L., Cai, L., Wang, X., & Yang, C. (2021). Single-Cell Sequencing Methodologies: From Transcriptome to Multi-Dimensional Measurement. *Small Methods*, *5*, 2100111.

Chen, Z., Chen, L., & Zhang, W. (2017). Tools for genomic and transcriptomic analysis of microbes at single-cell level. *Frontiers in Microbiology*, *8*(SEP).

Chung, M., Adkins, R. S., Mattick, J. S. A., Bradwell, K. R., Shetty, A. C., Sadzewicz, L., Tallon, L. J., Fraser, C. M., Rasko, D. A., Mahurkar, A., & Dunning Hotopp, J. C. (2021). FADU: a Quantification Tool for Prokaryotic Transcriptomic Analyses. *MSystems*, *6*(1).

Clingenpeel, S., Schwientek, P., Hugenholtz, P., & Woyke, T. (2014). Effects of sample treatments on genome recovery via single-cell genomics. *The ISME Journal*, *8*(12), 2546–2549.

Collins, D. J., Neild, A., deMello, A., Liu, A. Q., & Ai, Y. (2015). The Poisson distribution and beyond: methods for microfluidic droplet production and single cell encapsulation. *Lab on a Chip*, *15*(17), 3439–3459.

Dam, H. T., & Häggblom, M. M. (2017). Impact of estuarine gradients on reductive dechlorination of 1,2,3,4-tetrachlorodibenzo-p-dioxin in river sediment enrichment cultures. *Chemosphere*, *168*. https://doi.org/10.1016/j.chemosphere.2016.10.082

Dam, Hang T., Vollmers, J., Sobol, M. S., Cabezas, A., & Kaster, A. K. (2020). Targeted Cell Sorting Combined With Single Cell Genomics Captures Low Abundant Microbial Dark Matter With Higher Sensitivity Than Metagenomics. *Frontiers in Microbiology*, *11*, 1377.

De Anda, V., Chen, L. X., Dombrowski, N., Hua, Z. S., Jiang, H. C., Banfield, J. F., Li, W. J., & Baker, B. J. (2021). Brockarchaeota, a novel archaeal phylum with unique and versatile carbon cycling pathways. *Nature Communications 2021 12:1*, *12*(1), 1–12.

De Bourcy, C. F. A., De Vlaminck, I., Kanbar, J. N., Wang, J., Gawad, C., & Quake, S. R. (2014). A quantitative comparison of single-cell whole genome amplification methods. *PloS One*,

de Jager, V., & Siezen, R. J. (2011). Single-cell genomics: Unravelling the genomes of unculturable microorganisms. *Microbial Biotechnology*, *4*(4), 431–437.

Dean, F. B., Nelson, J. R., Giesler, T. L., & Lasken, R. S. (2001). Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Research*, *11*(6), 1095–1099.

Del Campo, C., Bartholomäus, A., Fedyunin, I., & Ignatova, Z. (2015). Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genetics*, *11*(10).

Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T., Rappé, M. S., MacLellan, S. L., Lücker, S., & Eren, A. M. (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology 2018 3:7*, *3*(7), 804–813.

Desai, C., Pathak, H., & Madamwar, D. (2010). Advances in molecular and "-omics" technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites. *Bioresource Technology*, *101*(6), 1558–1569.

DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319.

Díaz-Acosta, A., Sandoval, M. L., Delgado-Olivares, L., & Membrillo-Hernández, J. (2006). Effect of anaerobic and stationary phase growth conditions on the heat shock and oxidative stress responses in Escherichia coli K-12. *Archives of Microbiology*, *185*(6), 429–438.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, *10*(8), R85.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Dorfman, R. (1979). A Formula for the Gini Coefficient. *The Review of Economics and Statistics*, *61*(1), 146.

Doud, D. F. R., Bowers, R. M., Schulz, F., De Raad, M., Deng, K., Tarver, A., Glasgow, E., Vander Meulen, K., Fox, B., Deutsch, S., Yoshikuni, Y., Northen, T., Hedlund, B. P., Singer, S. W., Ivanova, N., & Woyke, T. (2019). Function-driven single-cell genomics uncovers cellulose-degrading bacteria from the rare biosphere. *The ISME Journal*, *14*(3), 659–675.

Doud, D. F. R., & Woyke, T. (2017). Novel approaches in function-driven single-cell genomics. *FEMS Microbiology Reviews*, *41*(4), 538–548.

Esteban, J. A., Salas, M., & Blanco, L. (1993). Fidelity of ф29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *The Journal of Biological Chemistry*, *268*(4), 2719–2726.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* , *32*(19), 3047–3048.

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, *320*(5879), 1034–1039.

Feder, M. E., & Hofmann, G. E. (1999). *Heat-shock proteins, molecular chaperones, and the stress response: Evolutionary and ecological physiology*.

Feng, W., Ueda, E., & Levkin, P. A. (2018). Droplet Microarrays: From Surface Patterning to High-Throughput Applications. *Advanced Materials* , *30*(20), 1706111.

Gadkar, V. J., & Filion, M. (2012). A Linear Concatenation Strategy to Construct 5 0-Enriched Amplified cDNA Libraries Using Multiple Displacement Amplification. *Springer*. https://doi.org/10.1007/s12033-012-9594-8

Gaillard, C., & Strauss, F. (1998). Avoiding adsorption of DNA to polypropylene tubes and denaturation of short DNA fragments. *Technical Tips Online*, *3*(1), 63–65.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E. J., Jayasinghe, L., Wright, C., … Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, *15*(3), 201–206.

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nature Reviews. Genetics*, *17*(3), 175–188.

Gich, F., Garcia-Gil, J., & Overmann, J. (2001). Previously unknown and phylogenetically diverse members of the green nonsulfur bacteria are indigenous to freshwater lakes. *Archives of Microbiology*, *177*(1), 1–10.

Goldstein, L. D., Chen, Y. J. J., Dunne, J., Mir, A., Hubschle, H., Guillory, J., Yuan, W., Zhang, J., Stinson, J., Jaiswal, B., Pahuja, K. B., Mann, I., Schaal, T., Chan, L., Anandakrishnan, S., Lin, C. W., Espinoza, P., Husain, S., Shapiro, H., … Modrusan, Z. (2017). Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics*, *18*(1), 1–10.

Gole, J., Gore, A., Richards, A., Chiu, Y. J., Fung, H. L., Bushman, D., Chiang, H. I., Chun, J., Lo, Y. H., & Zhang, K. (2013). Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nature Biotechnology*, *31*(12), 1126–1132.

Golonka, R., Yeoh, B. S., & Vijay-Kumar, M. (2019). The Iron Tug-of-War between Bacterial Siderophores and Innate Immunity. *Journal of Innate Immunity*, *11*(3), 249–262.

González-Cabaleiro, R., Mitchell, A. M., Smith, W., Wipat, A., & Ofiteru, I. D. (2017). Heterogeneity in pure microbial systems: Experimental measurements and modeling. *Frontiers in Microbiology*, *8*(SEP), 1813.

Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., Shaner, B., Annu, K., Putnam, D., Chen, W., Connelly, J., Pruett-Miller, S., Chen, X., Easton, J., & Gawad, C. (2021). Accurate genomic variant detection in single cells with primary template-directed amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(24), e2024176118.

Griffiths, B. S., Kuan, H. L., Ritz, K., Glover, L. A., McCaig, A. E., & Fenwick, C. (2004). The Relationship between Microbial Community Structure and Functional Stability, Tested Experimentally in an Upland Pasture Soil. *Microbial Ecology*, *47*(1), 104–113.

Griffiths, R. I., Whiteley, A. S., O'Donnell, A. G., & Bailey, M. J. (2000). Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Applied and Environmental Microbiology*, *66*(12), 5488–5491.

Grindberg, R. V., Ishoey, T., Brinza, D., Esquenazi, E., Coates, R. C., Liu, W. T., Gerwick, L., Dorrestein, P. C., Pevzner, P., Lasken, R., & Gerwick, W. H. (2011). Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PloS One*, *6*(4).

Gross, A., Schöndube, J., Niekrawitz, S., Streule, W., Riegger, L., Zengerle, R., & Koltay, P. (2013). Single-cell printer: automated, on demand, and label free. *Journal of Laboratory Automation*, *18*(6), 504–518.

Grünberger, F., Ferreira-Cerca, S., & Grohmann, D. (2022). Nanopore sequencing of RNA and cDNA molecules in Escherichia coli. *RNA* , *28*(3), 400–417.

Gu, W., Crawford, E. D., O'Donovan, B. D., Wilson, M. R., Chow, E. D., Retallack, H., & DeRisi, J. L. (2016). Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biology*, *17*(1), 1–13.

Gunasekera, T. S., Csonka, L. N., & Paliy, O. (2008). Genome-wide transcriptional responses of Escherichia coli K-12 to continuous osmotic and heat stresses. *Journal of Bacteriology*, *190*(10), 3712–3720.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* , *29*(8), 1072–1075.

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., & Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics*, *13*(1), 734.

Hall, E. W., Kim, S., Appadoo, V., & Zare, R. N. (2013). Lysis of a single cyanobacterium for whole genome amplification. *Micromachines*, *4*(3), 321–332.

Hanaki, K., Nakatake, H., Yamamoto, K., Odawara, T., & Yoshikura, H. (2000). DNase I activity retained after heat inactivation in standard buffer. *BioTechniques*, *29*(1), 38–42.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249.

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, *14*(1), 1–15.

Haroon, M. F., Skennerton, C. T., Steen, J. A., Lachner, N., Hugenholtz, P., & Tyson, G. W. (2013). *In-solution fluorescence in situ hybridization and Fluorescence-activated cell sorting for* *single cell and population genome recovery* (1st ed., Vol. 531, pp. 3–19). Elsevier Inc.

Hatzenpichler, R., Connon, S. A., Goudeau, D., Malmstrom, R. R., Woyke, T., & Orphan, V. J. (2016). Visualizing in situ translational activity for identifying and sorting slow-growing archaeal - bacterial consortia. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), E4069–E4078.

He, J., Du, S., Tan, X., Arefin, A., & Han, C. S. (2016). Improved lysis of single bacterial cells by a modified alkaline-thermal shock procedure. *BioTechniques*, *60*(3), 129–135.

Hedlund, B. P., Dodsworth, J. A., Murugapiran, S. K., Rinke, C., & Woyke, T. (2014). Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter." *Extremophiles: Life under Extreme Conditions*, *18*(5), 865–875.

Hemmerling, F., & Piel, J. (2022). Strategies to access biosynthetic novelty in bacterial genomes for drug discovery. *Nature Reviews Drug Discovery 2022 21:5*, *21*(5), 359–378.

Huber, J. A., Morrison, H. G., Huse, S. M., Neal, P. R., Sogin, M. L., & Mark Welch, D. B. (2009). Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environmental Microbiology*, *11*(5), 1292–1302.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., & Ise, K. (2016). A new view of the tree of life. *Nature Microbiology*, *1*, 16048.

Hug, L. A., Castelle, C. J., Wrighton, K. C., Thomas, B. C., Sharon, I., Frischkorn, K. R., Williams, K. H., Tringe, S. G., & Banfield, J. F. (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome*, *1*(1), 22.

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, *180*(18), 4765–4774.

Hutchison, C. A., Smith, H. O., Pfannkoch, C., & Venter, J. C. (2005). *Cell-free cloning using φ29 DNA polymerase*. *102*(48), 17332–17336.

Imdahl, F., & Saliba, A. E. (2020). Advances and challenges in single-cell RNA-seq of microbial communities. *Current Opinion in Microbiology*, *57*, 102–110.

Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A. E., & Vogel, J. (2020). Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nature Microbiology*, *5*(10), 1202–1206.

Islam, Z. F., Cordero, P. R. F., Feng, J., Chen, Y.-J., Bay, S. K., Jirapanjawat, T., Gleadow, R. M., Carere, C. R., Stott, M. B., Chiri, E., & Greening, C. (2019). Two Chloroflexi classes independently evolved the ability to persist on atmospheric hydrogen and carbon monoxide. *The ISME Journal*, *13*(7).

Jogia, G., Tronser, T., Popova, A., & Levkin, P. (2016). Droplet Microarray Based on Superhydrophobic-Superhydrophilic Patterns for Single Cell Analysis. *Microarrays*, *5*(4), 28.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Research*, *49*(D1), D545–D551.

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, *3*, e1165.

Kang, Y., McMillan, I., Norris, M. H., & Hoang, T. T. (2015). Single prokaryotic cell isolation and total transcript amplification protocol for transcriptomic analysis. *Nature Protocols*, *10*(7), 974–984.

Kang, Y., Norris, M. H., Zarzycki-Siek, J., Nierman, W. C., Donachie, S. P., & Hoang, T. T. (2011). Transcript amplification from single bacterium for transcriptome analysis. *Genome Research*, *21*(6), 925–935.

Kanter, I., & Kalisky, T. (2015). Single cell transcriptomics: Methods and applications. *Frontiers in Oncology*, *5*(FEB).

Kaster, A. K., Mayer-Blackwell, K., Pasarelli, B., & Spormann, A. M. (2014). Single cell genomic study of *Dehalococcoidetes* species from deep-sea sediments of the peruvian margin. *The ISME Journal*, *8*(9), 1831–1842.

Kaster, A. K., & Sobol, M. S. (2020). Microbial single-cell omics: the crux of the matter. *Applied Microbiology and Biotechnology*, *104*(19), 8209–8220.

Katz, L., & Baltz, R. H. (2016). Natural product discovery: past, present, and future. *Journal of Industrial Microbiology & Biotechnology*, *43*(2–3), 155–176.

Kogawa, M., Miyaoka, R., Hemmerling, F., Ando, M., Yura, K., Ide, K., Nishikawa, Y., Hosokawa, M., Ise, Y., Cahn, J. K. B., Takada, K., Matsunaga, S., Mori, T., Piel, J., Takeyama, H., & Nelson, K. E. (2022). Single-cell metabolite detection and genomics reveals uncultivated talented producer. *PNAS Nexus*, *1*(1), 1–13.

Köpke, B., Wilms, R., Engelen, B., Cypionka, H., & Sass, H. (2005). Microbial diversity in coastal subsurface sediments: a cultivation approach using various electron acceptors and substrate gradients. *Applied and Environmental Microbiology*, *71*(12), 7819–7830.

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* , *28*(24), 3211–3217.

Korfhage, C., Fricke, E., & Meier, A. (2015). Parallel WGA and WTA for comparative genome and transcriptome NGS analysis using tiny cell numbers. *Et al Current Protocols in Molecular Biology]*, *111*(1), 7.19.1-7.19.18.

Krueger, F., James, F., Ewels, P., Afyounian, E., & Schuster-Boeckler, B. (2021). *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*.

Kuchina, A., Brettner, L. M., Paleologu, L., Roco, C. M., Rosenberg, A. B., Carignano, A., Kibler, R., Hirano, M., DePaolo, R. W., & Seelig, G. (2021). Microbial single-cell RNA sequencing by split-pool barcoding. *Science*, *371*(6531).

Kumar, R., & Kumar, P. (2017). Future Microbial Applications for Bioenergy Production: A Perspective. *Frontiers in Microbiology*, *8*(MAR), 450.

Kuznetsova, I. M., Turoverov, K. K., & Uversky, V. N. (2014). What macromolecular crowding can do to a protein. In *International Journal of Molecular Sciences* (Vol. 15, Issue 12, pp. 23090–23140). Multidisciplinary Digital Publishing Institute.

Lan, F., Demaree, B., Ahmed, N., & Abate, A. R. (2017). Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nature Biotechnology*, *35*(7), 640–646.

Landry, Z., Swan, B. K., Herndl, G. J., Stepanauskas, R., & Giovannoni, S. J. (2017). SAR202 Genomes from the Dark Ocean Predict Pathways for the Oxidation of Recalcitrant Dissolved Organic Matter. *MBio*, *8*(2), e00413-17.

Lasken, R. S. (2009). Genomic DNA amplification by the multiple displacement amplification

(MDA) method. *Biochemical Society Transactions*, *37*(Pt 2), 450–453.

Lasken, R. S. (2013). Single-cell sequencing in its prime. *Nature Biotechnology 2013 31:3*, *31*(3), 211–212.

Lasken, R. S., & Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnology*, *7*.

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*, *12*(1), 124.

León-Zayas, R., Peoples, L., Biddle, J. F., Podell, S., Novotny, M., Cameron, J., Lasken, R. S., & Bartlett, D. H. (2017). The metabolic potential of the single cell genomes obtained from the Challenger Deep, Mariana Trench within the candidate superphylum Parcubacteria (OD1). *Environmental Microbiology*, *19*(7), 2769–2784.

Leung, K., Klaus, A., Lin, B. K., Laks, E., Biele, J., Lai, D., Bashashati, A., Huang, Y.-F. F., Aniba, R., Moksa, M., Steif, A., Mes-Masson, A.-M. M., Hirst, M., Shah, S. P., Aparicio, S., & Hansen, C. L. (2016). Robust high-performance nanoliter-volume single-cell multiple displacement amplification on planar substrates. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(30), 8484–8489.

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., & Ettema, T. J. G. (2020). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology 2020 19:4*, *19*(4), 225–240.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100.

Li, X., Zhang, P., Wang, H., & Yu, Y. (2022). Genes expressed at low levels raise false discovery rates in RNA samples contaminated with genomic DNA. *BMC Genomics*, *23*(1), 1–15.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* , *30*(7), 923–930.

Liu, Y., Jeraldo, P., Jang, J. S., Eckloff, B., Jen, J., & Walther-Antonio, M. (2019). Bacterial Single Cell Whole Transcriptome Amplification in Microfluidic Platform Shows Putative Gene Expression Heterogeneity. *Analytical Chemistry*, *91*(13), 8036–8044.

Liu, Y., Schulze-Makuch, D., de Vera, J.-P., Cockell, C., Leya, T., Baqué, M., Walther-Antonio, M., Liu, Y., Schulze-Makuch, D., de Vera, J.-P., Cockell, C., Leya, T., Baqué, M., & Walther-Antonio, M. (2018). The Development of an Effective Bacterial Single-Cell Lysis Method Suitable for Whole Genome Amplification in Microfluidic Platforms. *Micromachines*, *9*(8), 367.

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., & Crosby, L. (2018). Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *MSystems*, *3*(5).

Loffler, F. E., Yan, J., Ritalahti, K. M., Adrian, L., Edwards, E. A., Konstantinidis, K. T., Muller, J. A., Fullerton, H., Zinder, S. H., & Spormann, A. M. (2013). *Dehalococcoides mccarty*i gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, *Dehalococcoidia* classis nov., order *Dehalococcoidale*s ord. nov. and family *Dehalococcoidaceae* fam. nov., within the phylum Chloroflexi. *International Journal of Systematic and Evolutionary Microbiology*, *63*(Pt 2), 625–635.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A. R., Zhu, P., Hu, X., Xu, L., Yan, L., Bai, F., Qiao, J., Tang, F., Li, R., & Xie, X. S. (2012). Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science*, *338*(6114), 1627–1630.

Lv, H., Hung, C. S., & Henderson, J. P. (2014). Metabolomic analysis of siderophore cheater mutants reveals metabolic costs of expression in uropathogenic *Escherichia coli*. *Journal of Proteome Research*, *13*(3), 1397–1404.

Macaulay, I. C., Ponting, C. P., & Voet, T. (2017). Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics: TIG*, *33*(2), 155–168.

Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* , *27*(21), 2957–2963.

Maguire, F., Jia, B., Gray, K. L., Yin Venus Lau, W., Beiko, R. G., & L Brinkman, F. S. (2020). Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microbial Genomics*, *6*.

Manz, W., Amann, R., Ludwig, W., Vancanneyt, M., & Schleifer, K.-H. (1996). Application of a suite of 16S rRNA-specific oligonucleotide probes designed to investigate bacteria of the phylum cytophaga-flavobacter-bacteroides in the natural environment. *Microbiology (Reading, England)*, *142 (Pt 5)*(5), 1097–1106.

Marcy, Y., Ishoey, T., Lasken, R. S., Stockwell, T. B., Walenz, B. P., Halpern, A. L., Beeson, K. Y., Goldberg, S. M. D., & Quake, S. R. (2007). Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genetics*, *3*(9), 1702–1708.

Marcy, Y., Ouverney, C., Bik, E. M., Lösekann, T., Ivanova, N., Martin, H. G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D. A., & Quake, S. R. (2007). Dissecting biological dark matter with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(29), 11889–11894.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10.

Martins, B. M. C., & Locke, J. C. W. (2015). Microbial individuality: how single-cell heterogeneity enables population level strategies. *Current Opinion in Microbiology*, *24*, 104–112.

May, T., & Okabe, S. (2011). Enterobactin is required for biofilm development in reduced-genome Escherichia coli. *Environmental Microbiology*, *13*(12), 3149–3162.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618.

McKenzie, R. E., Keizer, E. M., Vink, J. N. A., van Lopik, J., Büke, F., Kalkman, V., Fleck, C., Tans, S. J., & Brouns, S. J. J. (2022). Single cell variability of CRISPR-Cas interference and adaptation. *Molecular Systems Biology*, *18*(4), e10680.

McLean, J. S., Lombardo, M. J., Badger, J. H., Edlund, A., Novotny, M., Yee-Greenbaum, J., Vyahhi, N., Hall, A. P., Yang, Y., Dupont, C. L., Ziegler, M. G., Chitsaz, H., Allen, A. E., Yooseph, S., Tesler, G., Pevzner, P. A., Friedman, R. M., Nealson, K. H., Venter, J. C., & Lasken, R. S. (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of*

Sciences of the United States of America, *110*(26), E2390–E2399.

Meier, H., Amann, R., Ludwig, W., & Schleifer, K. H. (1999). Specific oligonucleotide probes for in situ detection of a major group of gram-positive bacteria with low DNA G+C content. *Systematic and Applied Microbiology*, *22*(2), 186–196.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* , *34*(13), i142–i150.

Minton, A. P. (2001). The Influence of Macromolecular Crowding and Macromolecular Confinement on Biochemical Reactions in Physiological Media. *The Journal of Biological Chemistry*, *276*(14), 10577–10580.

Mollet, M., Godoy-Silva, R., Berdugo, C., & Chalmers, J. J. (2008). Computer simulations of the energy dissipation rate in a fluorescence-activated cell sorter: Implications to cells. *Biotechnology and Bioengineering*, *100*(2), 260–272.

Morawska, L. P., Hernandez-Valdes, J. A., & Kuipers, O. P. (2022). Diversity of bet-hedging strategies in microbial communities—Recent cases and insights. *WIREs Mechanisms of Disease*, *14*(2). https://doi.org/10.1002/wsbm.1544

Mosse, K. P. M., Patti, A. F., Smernik, R. J., Christen, E. W., & Cavagnaro, T. R. (2012). Physicochemical and microbiological effects of long- and short-term winery wastewater application to soils. *Journal of Hazardous Materials*, *201–202*, 219–228.

Mridha, S., & Kümmerli, R. (2021). From heterogeneity to homogeneity: coordination of siderophore gene expression among clonal cells of the bacterium Pseudomonas aeruginosa. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/2021.01.29.428812

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H. Y., Mojica, A., Chen, I.-M. A., Kyrpides, N. C., & Reddy, T. B. K. (2019). Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Research*, *47*(D1), D649–D659.

Müller, S., & Nebe-Von-Caron, G. (2010). Functional single-cell analyses: Flow cytometry and cell sorting of microbial populations and communities. *FEMS Microbiology Reviews*, *34*(4), 554–587.

Mullis, M. M., Rambo, I. M., Baker, B. J., & Reese, B. K. (2019). Diversity, Ecology, and Prevalence of Antimicrobials in Nature. *Frontiers in Microbiology*, *10*, 2518.

Murphy, C. L., Biggerstaff, J., Eichhorn, A., Ewing, E., Shahan, R., Soriano, D., Stewart, S., VanMol, K., Walker, R., Walters, P., Elshahed, M. S., & Youssef, N. H. (2021). Genomic characterization of three novel *Desulfobacterota* classes expand the metabolic and phylogenetic diversity of the phylum. *Environmental Microbiology*, *23*(8), 4326–4343.

Ngan, J. Y. G., Pasunooti, S., Tse, W., Meng, W., Ngan, S. F. C., Jia, H., Lin, J. Q., Ng, S. W., Jaafa, M. T., Cho, S. L. S., Lim, J., Koh, H. Q. V., Ghani, N. A., Pethe, K., Sze, S. K., Lescar, J., & Alonso, S. (2021). HflX is a GTPase that controls hypoxia-induced replication arrest in slow-growing mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(12), e2006717118.

Nishikawa, Y., Hosokawa, M., Maruyama, T., Yamagishi, K., Mori, T., & Takeyama, H. (2015). Monodisperse picoliter droplets for low-bias and contamination-free reactions in single-cell whole genome amplification. *PloS One*, *10*(9).

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel, S. R., Woyke, T., Mclean, J. S., Lasken, R., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2013). Assembling Single-Cell Genomes and Mini-Metagenomes From Chimeric MDA Products. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *20*(10), 714–737.

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, *4*(1), 14.

Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., Poulton, N. J., Burkart, M. D., La Clair, J. J., Chisholm, S. W., & Stepanauskas, R. (2019). Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell*, *179*(7), 1623-1635.e11.

Paez, J. G., Lin, M., Beroukhim, R., Lee, J. C., Zhao, X., Richter, D. J., Gabriel, S., Herman, P., Sasaki, H., Altshuler, D., Li, C., Meyerson, M., & Sellers, W. R. (2004). Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Research*, *32*(9), e71–e71.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, *6*(1), 1–11.

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P. A., & Hugenholtz, P. (2022). GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785–D794.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055.

Pati, A., LaButti, K., Pukall, R., Nolan, M., Rio, T. G. D., Tice, H., Cheng, J.-F., Lucas, S., Chen, F., Copeland, A., Ivanova, N., Mavromatis, K., Mikhailova, N., Pitluck, S., Bruce, D., Goodwin, L., Land, M., Hauser, L., Chang, Y.-J., … Lapidus, A. (2010). Complete genome sequence of Sphaerobacter thermophilus type strain (S 6022T). *Standards in Genomic Sciences*, *2*(1), 49.

Peralta, D. R., Adler, C., Corbalán, N. S., Paz García, E. C., Pomares, M. F., & Vincent, P. A. (2016). Enterobactin as Part of the Oxidative Stress Response Repertoire. *PloS One*, *11*(6), e0157799.

Pernthaler, J., Glöckner, F. O., Schönhuber, W., & Amann, R. (2001). Fluorescence in situ hybridization with rRNA-targeted oligonucleotide probes. *Methods in Microbiology*, *30*(3), 207–226.

Picelli, S. (2017). Single-cell RNA-sequencing: The future of genome biology is now. *RNA Biology*, *14*(5), 637–650.

Podar, M., Abulencia, C. B., Walcher, M., Hutchison, D., Zengler, K., Garcia, J. A., Holland, T., Cotton, D., Hauser, L., & Keller, M. (2007). Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Applied and Environmental Microbiology*, *73*(10), 3205–3214.

Pratscher, J., Vollmers, J., Wiegand, S., Dumont, M. G., & Kaster, A.-K. (2018). Unravelling the Identity, Metabolic Potential and Global Biogeography of the Atmospheric Methane-Oxidizing Upland Soil Cluster α. *Environmental Microbiology*, *20*(3), 1016–1029.

Prezza, G., Heckel, T., Dietrich, S., Homberger, C., Westermann, A. J., & Vogel, J. (2020). Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. *RNA* , *26*(8), 1069–1078.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods*, *8*(1), 12–24.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *70*(1), e102.

Probst, J., Brechtel, S., Scheel, B., Herr, I., Jung, G., Rammensee, H. G., & Pascolo, S. (2006). Characterization of the ribonuclease activity on the skin surface. *Genetic Vaccines and Therapy*, *4*, 4.

Pruesse, E., Peplies, J., & Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* , *28*(14), 1823–1829.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. R Foundation for Statistical Computing*.

Raghunathan, A., Ferguson, H. R., Bornarth, C. J., Song, W., Driscoll, M., & Lasken, R. S. (2005). Genomic DNA amplification from a single bacterium. *Applied and Environmental Microbiology*, *71*(6), 3342–3347.

Ralston, G. B. (1990). Effects of crowding in protein solutions. *Journal of Chemical Education*, *67*(10), 857–860.

Rezaei, M., Radfar, P., Winter, M., McClements, L., Thierry, B., & Warkiani, M. E. (2021). Simple-to-operate approach for single cell analysis using a hydrophobic surface and nanosized droplets. *Analytical Chemistry*, *93*(10), 4584–4592.

Rhee, M., Light, Y. K., Meagher, R. J., & Singh, A. K. (2016). Digital Droplet Multiple Displacement Amplification (ddMDA) for Whole Genome Sequencing of Limited DNA Samples. *PloS One*, *11*(5), e0153699.

Riba, J., Gleichmann, T., Zimmermann, S., Zengerle, R., & Koltay, P. (2016). Label-free isolation and deposition of single bacterial cells from heterogeneous samples for clonal culturing. *6*(1), 1–9.

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., Dmitrieff, E., Malmstrom, R., Stepanauskas, R., & Woyke, T. (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nature Protocols*, *9*(5), 1038–1048.

Rinke, C., Low, S., Woodcroft, B. J., Raina, J.-B., Skarshewski, A., Le, X. H., Butler, M. K., Stocker, R., Seymour, J., Tyson, G. W., & Hugenholtz, P. (2016). Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ*, *4*(e2486), e2486.

Roberfroid, S., Vanderleyden, J., & Steenackers, H. (2016). Gene expression variability in clonal populations: Causes and consequences. *Critical Reviews in Microbiology*, *42*(6), 969–984.

Rodrigue, S., Malmstrom, R. R., Berlin, A. M., Birren, B. W., Henn, M. R., & Chisholm, S. W. (2009). Whole Genome Amplification and De novo Assembly of Single Bacterial Cells. *PloS One*, *4*(9), e6864.

Ruan, Q., Ruan, W., Lin, X., Wang, Y., Zou, F., Zhou, L., Zhu, Z., & Yang, C. (2020). Digital-WGS: Automated, highly efficient whole-genome sequencing of single cells by digital microfluidics. *Science Advances*, *6*(50).

Sabina, J., & Leamon, J. H. (2015). Bias in whole genome amplification: Causes and considerations. In T. Kroneis (Ed.), *Methods in Molecular Biology* (Vol. 1347, pp. 15–41). Humana Press Inc.

Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., Funk, K., Ketter, A., Kim, S., Kimchi, A., Kitts, P. A., Kuznetsov, A., Lathrop, S., Lu, Z., McGarvey, K., Madden, T. L., Murphy, T. D., O'Leary, N., Phan, L., … Ostell, J. (2020). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *48*(D1), D9–D16.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* , *30*(14), 2068–2069.

Shakoori, A. R. (2017). Fluorescence In Situ hybridization (FISH) and its applications. In *Chromosome Structure and Aberrations* (pp. 343–367). Springer India.

Sidore, A. M., Lan, F., Lim, S. W., & Abate, A. R. (2015). Enhanced sequencing coverage with digital droplet multiple displacement amplification. *Nucleic Acids Research*, *44*(7), e66–e66.

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F.

(2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843.

Siguret, V., Ribba, A. S., Cherel, G., Meyer, D., & Pietu, G. (1994). Effect of plasmid size on transformation efficiency by electroporation of Escherichia coli DH5α. *BioTechniques*, *16*(3), 422–426.

Sim, M., Koirala, S., Picton, D., Strahl, H., Hoskisson, P. A., Rao, C. V., Gillespie, C. S., & Aldridge, P. D. (2017). Growth rate control of flagellar assembly in Escherichia coli strain RP437. *Scientific Reports 2017 7:1*, *7*(1), 1–11.

Skennerton, C. T., Imelfort, M., & Tyson, G. W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Research*, *41*(10), e105.

Sommer, D. D., Delcher, A. L., Salzberg, S. L., & Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*, *8*(1), 64.

Sorek, R., Kunin, V., & Hugenholtz, P. (2008). CRISPR - A widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews. Microbiology*, *6*(3), 181–186.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications 2021 12:1*, *12*(1), 1–15.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., & Delong, E. F. (1996). Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, *178*(3), 591–599.

Stepanauskas, R. (2012). Single cell genomics: An individual look at microbes. *Current Opinion in Microbiology*, *15*(5), 613–620.

Stepanauskas, R., Fergusson, E. A., Brown, J., Poulton, N. J., Tupper, B., Labonté, J. M., Becraft, E. D., Brown, J. M., Pachiadaki, M. G., Povilaitis, T., Thompson, B. P., Mascena, C. J., Bellows, W. K., & Lubys, A. (2017). Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nature Communications*, *8*(1), 84.

Stepanauskas, R., & Sieracki, M. E. (2007). Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(21), 9052–9057.

Stewart, E. J. (2012). Growing Unculturable Bacteria. *Journal of Bacteriology*, *194*(16), 4151.

Stincone, P., & Brandelli, A. (2020). Marine bacteria as source of antimicrobial compounds. *Critical Reviews in Biotechnology*, *40*(3), 306–319.

Swan, B. K., Martinez-Garcia, M., Preston, C. M., Sczyrba, A., Woyke, T., Lamy, D., Reinthaler, T., Poulton, N. J., Masland, E. D. P., Gomez, M. L., Sieracki, M. E., DeLong, E. F., Herndl, G. J., & Stepanauskas, R. (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* , *333*(6047), 1296–1300.

Tarazona, S., García, F., Ferrer, A., Dopazo, J., & Conesa, A. (2012). NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.Journal*, *17*(B), 18.

Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjöld, M., Ponder, B. A. J., & Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics*, *13*(3), 718–725.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43.

Van Rossum, T., Ferretti, P., Maistrenko, O. M., & Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology 2020 18:9*, *18*(9), 491–506.

Villanueva, R. A. M., & Chen, Z. J. (2019). ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 160–167.

Vollmers, J., Wiegand, S., & Kaster, A. K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! In *PLoS ONE* (Vol. 12, Issue 1, pp. 1–31).

Vollmers, J., Wiegand, S., Lenk, F., & Kaster, A.-K. (2022). How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Research*, *1*.

Wang, J., Chen, L., Chen, Z., & Zhang, W. (2015). RNA-seq based transcriptomic analysis of single bacterial cells. *Integrative Biology*, *7*(11), 1466–1476.

Wang, S., Deng, K., Zaremba, S., Deng, X., Lin, C., Wang, Q., Lou Tortorello, M., & Zhang, W. (2009). Transcriptomic response of *Escherichia coli* O157:H7 to oxidative stress. *Applied and Environmental Microbiology*, *75*(19), 6110–6123.

Wang, Y., & Navin, N. E. (2015). Advances and Applications of Single-Cell Sequencing Technologies. *Molecular Cell*, *58*(4), 598–609.

Westram, R., Bader, K., Pruesse, E., Kumar, Y., Meier, H., Gloeckner, F. O., & Ludwig, W. (2011). ARB: a software environment for sequence data. In F. J. de Bruijin (Ed.), *Handbook of molecular microbial ecology I: metagenomics and complementary approaches* (pp. 399–406). John Wiley & Sons, Inc.

West-Roberts, J. A., Matheus-Carnevali, P. B., Schoelmerich, M. C., Al-Shayeb, B., Thomas, A. D., Sharrar, A., He, C., Chen, L.-X., Lavy, A., Keren, R., Amano, Y., & Banfield, J. F. (2021). The Chloroflexi supergroup is metabolically diverse and representatives have novel genes for non-photosynthesis based CO2 fixation. *BioRxiv*, 2021.08.23.457424.

Wiegand, S., Dam, H. T., Riba, J., Vollmers, J., & Kaster, A. K. (2021). Printing Microbial Dark Matter: Using Single Cell Dispensing and Genomics to Investigate the Patescibacteria/Candidate Phyla Radiation. *Frontiers in Microbiology*, *12*, 1512.

Wiegand, S., Jogler, M., Boedeker, C., Pinto, D., Vollmers, J., Rivas-Marín, E., Kohn, T., Peeters, S. H., Heuer, A., Rast, P., Oberbeckmann, S., Bunk, B., Jeske, O., Meyerdierks, A., Storesund, J. E., Kallscheuer, N., Lücker, S., Lage, O. M., Pohl, T., … Jogler, C. (2020). Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. *Nature Microbiology*, *5*(1), 126–140.

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*, 1338.

Winter, J., Linke, K., Jatzek, A., & Jakob, U. (2005). Severe oxidative stress causes inactivation of

DnaK and activation of the redox-regulated chaperone Hsp33. *Molecular Cell*, *17*(3), 381–392.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090.

Wolfien, M., David, R., & Galow, A.-M. (2021). Single-Cell RNA Sequencing Procedures and Data Analysis. In *Bioinformatics* (pp. 19–35). Exon Publications.

Wongsurawat, T., Jenjaroenpun, P., Taylor, M. K., Lee, J., Tolardo, A. L., Parvathareddy, J., Kandel, S., Wadley, T. D., Kaewnapan, B., Athipanyasilp, N., Skidmore, A., Chung, D., Chaimayo, C., Whitt, M., Kantakamalakul, W., Sutthent, R., Horthongkham, N., Ussery, D. W., Jonsson, C. B., & Nookaew, I. (2019). Rapid Sequencing of Multiple RNA Viruses in Their Native Form. *Frontiers in Microbiology*, *10*(FEB), 260.

Woodcroft, B., Lamberton, T., & Imelfort, M. (2019). *BamM: Metagenomics-focused BAM file manipulation*. https://github.com/ecogenomics/BamM

Woyke, T., Xie, G., Copeland, A., González, J. M., & Han, C. (2009). Assembling the Marine Metagenome, One Cell at a Time. *PloS One*, *4*(4), 5299.

Woyke, Tanja, Doud, D. F. R., & Schulz, F. (2017). The trajectory of microbial single-cell sequencing. *Nature Methods*, *14*(11), 1045–1054.

Woyke, Tanja, Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R., & Cheng, J.-F. (2011). Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. *PloS One*, *6*(10), e26161.

Woyke, Tanja, Tighe, D., Mavromatis, K., Clum, A., Copeland, A., Schackwitz, W., Lapidus, A., Wu, D., Mccutcheon, J. P., Mcdonald, B. R., Moran, N. A., Bristow, J., & Cheng, J. F. (2010). One bacterial cell, one complete genome. *PloS One*, *5*(4).

Wu, D., Raymond, J., Wu, M., Chatterji, S., Ren, Q., Graham, J. E., Bryant, D. A., Robb, F., Colman, A., Tallon, L. J., Badger, J. H., Madupu, R., Ward, N. L., & Eisen, J. A. (2009). Complete genome sequence of the aerobic CO-oxidizing thermophile Thermomicrobium roseum. *PloS One*, *4*(1).

Wu, L., Liu, X., Schadt, C. W., & Zhou, J. (2006). Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. *Applied and Environmental Microbiology*, *72*(7), 4931–4941.

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* , *32*(4), 605–607.

Xu, L., Brito, I. L., Alm, E. J., & Blainey, P. C. (2016). Virtual microfluidics for digital quantification and single-cell sequencing. *Nature Methods*, *13*(9), 759–762.

Yanling Song, Xing Xu, Wei Wang, Tian Tian, Zhi Zhu, & Chaoyong Yang. (2019). Single cell transcriptomics: moving towards multi-omics. *The Analyst*, *144*(10), 3172–3189.

Yilmaz, S., Haroon, M. F., Rabkin, B. A., Tyson, G. W., & Hugenholtz, P. (2010). Fixation-free fluorescence in situ hybridization for targeted enrichment of microbial populations. *The ISME Journal*, *4*(10), 1352–1356.

Yoon, S. H., Han, M.-J., Lee, S. Y., Jeong, K. J., & Yoo, J.-S. (2003). Combined transcriptome and proteome analysis of Escherichia coli during high cell density culture. *Biotechnology and Bioengineering*, *81*(7), 753–767.

Yosef, I., Goren, M. G., Kiro, R., Edgar, R., & Qimron, U. (2011). High-temperature protein G is essential for activity of the *Escherichia coli* clustered regularly interspaced short palindromic repeats (CRISPR)/Cas system. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 20136–20141.

Zeileis, A., Kleiber, C., Rep., M. A. Z.-. T., & 2009, U. (2014). Package “ineq.” *Cran.Microsoft.Com*. https://cran.microsoft.com/snapshot/2014-09-08/web/packages/ineq/ineq.pdf

Zhang, C.-Z., Adalsteinsson, V. A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K. L., Meyerson, M., Love, J. C., & Author, N. C. (2016). *Calibrating genomic and allelic coverage bias in single-cell sequencing HHS Public Access Author manuscript*.

Zhang, D. Y., Brandwein, M., Hsuih, T., & Li, H. B. (2001). Ramification amplification: A novel isothermal DNA amplification method. *Molecular Diagnosis: A Journal Devoted to the Understanding of Human Disease through the Clinical Application of Molecular Biology*, *6*(2), 141–150.

Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W., & Church, G. M.

(2006). Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology*, *24*(6), 680–686.

Zhang, Y., Gao, J., Huang, Y., & Wang, J. (2018). Biophysical Perspective Recent Developments in Single-Cell RNA-Seq of Microorganisms. *Biophysj*, *115*, 173–180.

Zheng, Y., Saitou, A., Wang, C.-M., Toyoda, A., Minakuchi, Y., Sekiguchi, Y., Ueda, K., Takano, H., Sakai, Y., Abe, K., Yokota, A., & Yabe, S. (2019). Genome Features and Secondary Metabolites Biosynthetic Potential of the Class *Ktedonobacteria*. *Frontiers in Microbiology*, *10*, 893.

Zimmerman, S. B., & Harrison, B. (1987). Macromolecular crowding increases binding of DNA polymerase to DNA: an adaptive effect. *Proceedings of the National Academy of Sciences*, *84*(7), 1871–1875.

Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, *338*(6114), 1622–1626.

6 Appendix

Appendix Table 1. MDA sample summary
Cq= quantification cycle number, the cycle number at which the fluorescence first rises above the threshold level. RFU= relative fluorescence units.

Sample Name	Cq	RFU Endpoint	Total DNA (ng)
0.8 µL_1	49.15	693	20.896
0.8 µL_2	52.82	432	5.088
0.8 µL_3	45.18	748	48.96
0.8 µL_4	47.39	782	54.08
0.8 µL_5	41.38	3353	46.4
Std Dev	*4.2829*	*1211*	*21.1098*
Average	*47.184*	*1202*	*35.0848*
1.0 µL_1	38.05	4087	59.52
1.0 µL_2	44.91	3315	91.52
1.0 µL_3	44.48	3749	35.2
1.0 µL_4	41.26	3223	64.96
1.0 µL_5	50.8	692	12.13
Std Dev	*4.7476*	*1343*	*30.2324*
Average	*43.9*	*3013*	*52.6660*
1.25 µL_1	34.94	6976	83.25
1.25 µL_2	30.59	6913	125.55
1.25 µL_3	35.49	6695	63.45
1.25 µL_4	22.35	12238	707.2
1.25 µL_5	24.09	11769	611.2
Std Dev	*6.0633*	*2823*	*313.9996*
Average	*29.492*	*8918*	*318.1300*
5 µL_1	45.29	45308	1,190.40
5 µL_2	43.56	41385	1,670.40
5 µL_3	42.48	53918	1,772.80
5 µL_4	40.16	61054	992
5 µL_5	36.7	61175	1,017.60
Std Dev	*3.3282*	*9023*	*368.5311*
Average	*41.638*	*52568*	*1328.6400*
10 µL_1	34.12	59245	1,369.60
10 µL_2	43.17	52477	1,881.60
10 µL_3	37.06	58636	2,348.80
10 µL_4	42.17	51944	2,515.20
10 µL_5	39.26	52845	2,515.20
Std Dev	*3.7068*	*3591*	*496.1012*
Average	*39.156*	*55029*	*2126.0800*

Appendix Table 2. MDA read processing summary
Total reads from all samples were used as input for read trimming in Trim Galore. After read trimming, all samples were down-sampled to 200X sequence depth prior to decontamination with FASTQ-Screen and duplicate removal with dedupe.sh from BBTools. P-value was calculated Anova: Single Factor with an alpha value of 0.05 in Microsoft Excel®.

Sample Name	Reads Lost from Trimming	Non-contaminated Reads	PCR Duplicates
0.8 µL_1	0.13%	98.55%	1.58%
0.8 µL_2	0.12%	98.38%	2.06%
0.8 µL_3	0.09%	98.84%	1.94%
0.8 µL_4	0.08%	91.11%	1.72%
0.8 µL_5	0.12%	98.51%	0.92%
Std Dev	*0.0002*	*0.0334*	*0.0045*
Average	*0.1064%*	*97.0765%*	*1.6440%*
1.0 µL_1	0.13%	97.89%	1.24%
1.0 µL_2	0.12%	95.35%	0.97%
1.0 µL_3	0.09%	99.44%	2.32%
1.0 µL_4	0.09%	94.75%	0.97%
1.0 µL_5	0.09%	98.50%	1.37%
Std Dev	*0.00019*	*0.0204*	*0.0056*
Average	*0.1047%*	*97.1864%*	*1.3740%*
1.25 µL_1	0.03%	96.52%	1.41%
1.25 µL_2	0.03%	96.26%	1.43%
1.25 µL_3	0.04%	95.74%	1.75%
1.25 µL_4	0.05%	97.57%	1.28%
1.25 µL_5	0.05%	96.60%	1.22%
Std Dev	*0.0001*	*0.0067*	*0.0021*
Average	*0.0396%*	*96.5391%*	*1.4180%*
5 µL_1	0.08%	93.65%	2.79%
5 µL_2	0.06%	98.25%	6.65%
5 µL_3	0.04%	98.03%	5.01%
5 µL_4	0.06%	19.33%	1.11%
5 µL_5	0.09%	34.24%	2.31%
Std Dev	*0.0002*	*0.3867*	*0.0223*
Average	*0.0665%*	*68.6991%*	*3.5740%*
10 µL_1	0.08%	55.66%	1.55%
10 µL_2	0.09%	96.40%	50.82%
10 µL_3	0.14%	72.44%	2.71%
10 µL_4	0.06%	97.09%	31.26%
10 µL_5	0.07%	4.37%	1.16%
Std Dev	*0.0003*	*0.3819*	*0.2258*
Average	*0.0887%*	*65.1911%*	*17.5000%*
p-value	*0.0002*	*0.0960*	*0.0870*

Appendix Table 3. MDA read mapping summary

After contaminant and duplicate removal, reads were mapped to *E. coli* MG1655 reference genome with *bbmap.sh*. Mapping statistics were calculated with BBmap as well, Gini indices were calculated with the ineq package in R studio. P-value was calculated Anova: Single Factor with an alpha value of 0.05 in Microsoft Excel®.

Sample Name	**Mapping Coverage (%)**	**Avg. Insert Size**	**Read Std dev/10 Kb**	**Gini Index**
0.8 µL_1	59.46	223.03	572.965	0.8708
0.8 µL_2	58.17	211.67	750.424	0.9022
0.8 µL_3	58.13	209.02	664.428	0.8836
0.8 µL_4	47.45	212.57	578.596	0.8907
0.8 µL_5	95.03	239.46	380.321	0.6733
Std Dev	*18.2003*	*12.5441*	*137.5307*	*0.0962*
Average	*63.6494*	*219.1500*	*589.3468*	*0.8441*
1.0 µL_1	69.73	231.81	598.19	0.8570
1.0 µL_2	85.01	231.36	459.222	0.7783
1.0 µL_3	21.20	212.69	927.816	0.9557
1.0 µL_4	67.73	226.17	417.53	0.8096
1.0 µL_5	53.86	226.44	653.862	0.8913
Std Dev	*24.0911*	*7.7362*	*201.7674*	*0.0696*
Average	*59.5071*	*225.6940*	*611.3240*	*0.8584*
1.25 µL_1	90.63	200.66	299.296	0.6762
1.25 µL_2	96.14	195.93	304.19	0.6360
1.25 µL_3	78.95	213.10	471.041	0.7786
1.25 µL_4	94.53	241.56	513.246	0.6831
1.25 µL_5	64.79	237.90	490.337	0.7967
Std Dev	*13.1510*	*20.9923*	*105.039*	*0.0698*
Average	*85.0079*	*217.8300*	*415.6220*	*0.7141*
5 µL_1	85.73	216.86	977.542	0.8520
5 µL_2	81.74	223	1179.344	0.8868
5 µL_3	82.06	235	1192.636	0.8862
5 µL_4	48.39	212.43	175.801	0.8889
5 µL_5	32.67	201.55	451.882	0.9372
Std Dev	*24.0625*	*12.4163*	*458.3374*	*0.0304*
Average	*66.1188*	*217.7680*	*795.4410*	*0.8902*
10 µL_1	61.45	224.95	385.936	0.8604
10 µL_2	24.69	255.58	714.113	0.9761
10 µL_3	62.34	219.01	862.061	0.9161
10 µL_4	53.45	211.13	818.291	0.9153
10 µL_5	23.72	238.46	71.785	0.9341
Std Dev	*19.4160*	*17.5222*	*335.3406*	*0.0106*
Average	*45.1300*	*229.8260*	*570.4372*	*0.9204*
p-value	0.0736	0.6265	0.3561	0.0009

Appendix Table 4. MDA assembly statistic summary

After read processing, the final sequence depth used for assembly was calculated. Assembly N50, length, and coverage were calculated using QUAST. Coverage is calculated as the percent of contigs aligned to the reference genome. MDMcleaner was used to calculate completeness and assembly contamination. P-value was calculated Anova: Single Factor with an alpha value of 0.05 in Microsoft Excel®.

Sample Name	Final Sequence Depth (X)	N50	Length (bp)	Assembly Coverage	Genome Completeness	Fraction of Untrusted Base-pairs
0.8 µL_1	45	37,203	2,008,434	43.18%	90%	0.21%
0.8 µL_2	40	31,827	1,826,589	39.29%	55%	0.50%
0.8 µL_3	44	28,055	2,033,477	43.71%	55%	0.03%
0.8 µL_4	38	31,357	1,508,083	32.39%	75%	0.48%
0.8 µL_5	87	51,543	4,023,801	86.59%	100%	0.18%
Std Dev	*20*	*9,289*	*997,096*	*21.48*	*20%*	*0.20%*
Average	*51*	*35,997*	*2,280,077*	*49.03*	*75%*	*0.28%*
1.0 µL_1	50	30,920	2,510,486	54.00%	85%	0.26%
1.0 µL_2	65	36,080	3,151,225	67.81%	95%	0.29%
1.0 µL_3	20	40,727	711,913	15.32%	55%	0.11%
1.0 µL_4	57	32,298	2,462,399	52.94%	80%	0.08%
1.0 µL_5	41	23,582	1,846,315	39.66%	70%	0.28%
Std Dev	*17*	*6,370*	*920,516*	*19.81*	*15%*	*0.10%*
Average	*46*	*32,721*	*2,136,468*	*45.94*	*77%*	*0.21%*
1.25 µL_1	89	56,529	3,630,701	78.16%	95%	0.23%
1.25 µL_2	100	67,482	4,156,954	89.46%	95%	0.21%
1.25 µL_3	71	39,389	3,179,043	68.40%	90%	0.34%
1.25 µL_4	82	26,847	4,034,821	86.81%	100%	0.19%
1.25 µL_5	61	40,650	2,612,738	56.20%	90%	0.19%
Std Dev	*15*	*15,901*	*636,856*	*13.71*	*4%*	*0.06%*
Average	*81*	*46,179*	*3,522,851*	*75.81*	*94%*	*0.23%*
5 µL_1	52	26,847	2,945,211	63.31%	95%	0.28%
5 µL_2	41	21,803	2,673,880	57.49%	90%	0.14%
5 µL_3	41	23,782	2,720,379	58.46%	90%	0.10%
5 µL_4	19	13,693	1,347,172	28.91%	55%	0.57%
5 µL_5	18	12,525	940,124	20.19%	40%	0.34%
Std Dev	*15*	*6,319*	*913,433*	*19.65*	*25%*	*0.19%*
Average	*34*	*19,730*	*2,125,353*	*45.67*	*74%*	*0.29%*
10 µL_1	36	22,962	1,973,202	42.25%	65%	0.47%
10 µL_2	5	15,114	494,824	10.64%	30%	1.62%
10 µL_3	31	19,415	1,768,861	37.97%	65%	0.75%
10 µL_4	28	16,327	1,568,571	33.69%	50%	0.51%
10 µL_5	5	12,760	550,195	11.82%	35%	0.08%
Std Dev	*14*	*3,964*	*698,482*	*14.96*	*16%*	*0.57%*
Average	*21*	*17,316*	*1,271,131*	*27.27*	*49%*	*0.68%*
p-value	*0.0002*	*0.0004*	*0.0088*	*0.0087*	*0.0128*	*0.0944*

Appendix Table 5. scWTA gene coverage summary

Gene coverage was calculated when a gene had >5 transcripts mapped per a sample.

Sample	Cell Number	Condition	Gene Coverage
E16April	1	heat-shock	10.2842
F15April	1	heat-shock	8.1796
G14April	1	heat-shock	4.3827
G15April	1	heat-shock	2.1480
F16April	1	heat-shock	8.8089
F17April	1	heat-shock	16.6414
G16April	1	heat-shock	5.0119
E14April	1	heat-shock	6.1619
E15April	1	heat-shock	6.5524
E17April	1	heat-shock	7.4203
F14April	1	heat-shock	5.3157
G17April	1	heat-shock	4.5780
G16Sep	1	heat-shock	11.8464
F18Sep	1	heat-shock	9.4598
G19Sep	1	heat-shock	10.2408
E17Sep	1	heat-shock	3.8620
E19Sep	1	heat-shock	4.3393
G17Sep	1	heat-shock	14.7754
G18Sep	1	heat-shock	10.3276
F19Sep	1	heat-shock	6.8345
G15Sep	1	heat-shock	6.0968
E9April	1	non-treated	6.1402
F10April	1	non-treated	3.8837
E10April	1	non-treated	6.7477
F8April	1	non-treated	11.9766
G8April	1	non-treated	6.4005
E8April	1	non-treated	6.6392
F9April	1	non-treated	4.1875
F11April	1	non-treated	5.1204
G9April	1	non-treated	12.4105
G10April	1	non-treated	7.9410
G11April	1	non-treated	5.7496
E9March	1	non-treated	4.1441
E10March	1	non-treated	6.0968
E11March	1	non-treated	4.0790
E12March	1	non-treated	5.4459
E13March	1	non-treated	3.7101
E14March	1	non-treated	5.5327
E15March	1	non-treated	4.6214
E09Sep	1	non-treated	11.7379
F07Sep	1	non-treated	6.5958
E07Sep	1	non-treated	8.3098
F08Sep	1	non-treated	9.1777
F11Sep	1	non-treated	6.6175
		average	**7.1939**

Sample	Cell Number	Condition	Gene Coverage
I15April	10	heat-shock	6.9212
I16April	10	heat-shock	6.1402
I14April	10	heat-shock	24.4088
I17Sep	10	heat-shock	3.8403
I8April	10	non-treated	19.6355
I9April	10	non-treated	9.8069
I11April	10	non-treated	9.9588
G9March	10	non-treated	12.9963
G10March	10	non-treated	8.5702
I10Sep	10	non-treated	3.6884
		average	**10.5967**
K14April	100	heat-shock	24.0616
K15April	100	heat-shock	30.8961
K16April	100	heat-shock	34.4109
K17Sep	100	heat-shock	12.8878
K9April	100	non-treated	56.4982
K8April	100	non-treated	57.4528
K11April	100	non-treated	30.5923
G13March	100	non-treated	22.6079
G15March	100	non-treated	58.6895
K10Sep	100	non-treated	45.0423
		average	37.3140
HS1Sep	Bulk	heat-shock	94.8145
HS2Sep	Bulk	heat-shock	82.2521
C1Sep	Bulk	non-treated	73.9206
C2Sep	Bulk	non-treated	78.6939
M17Sep	positive	heat-shock	21.3278
L13Sep	positive	heat-shock	8.1796
M09Sep	positive	non-treated	17.3356
M08Sep	positive	non-treated	22.8466
		average	**49.9213**

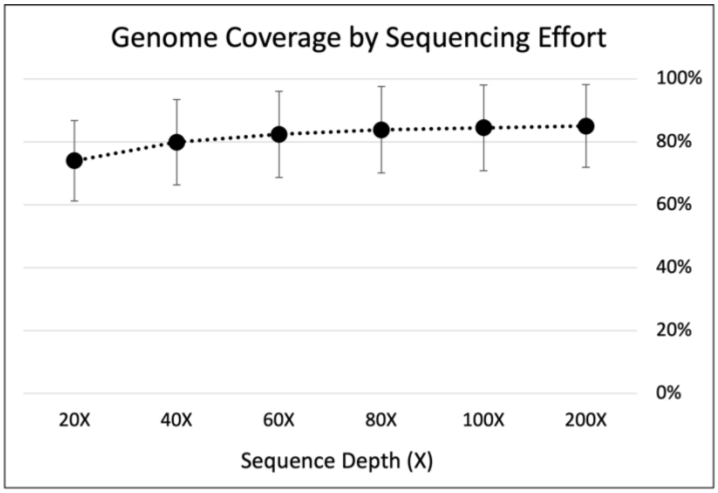
Appendix Table 6. DESeq2 result summary for top 8 transcribed heat-shock genes

Values were calculated based on differences solely between heat-shock and non-treated single-cells. A postive log2FoldChange means that expression in heat-shock cells is greater than non-treated cells. Bolded genes met the threshold (padj <0.05) for significant differential expression. baseMean is the mean normalized counts for all samples, lfcSE is the standard error, pvalue is the Wald test p-value, and padj is the Benjamini-Hochberg adjusted pvalue.

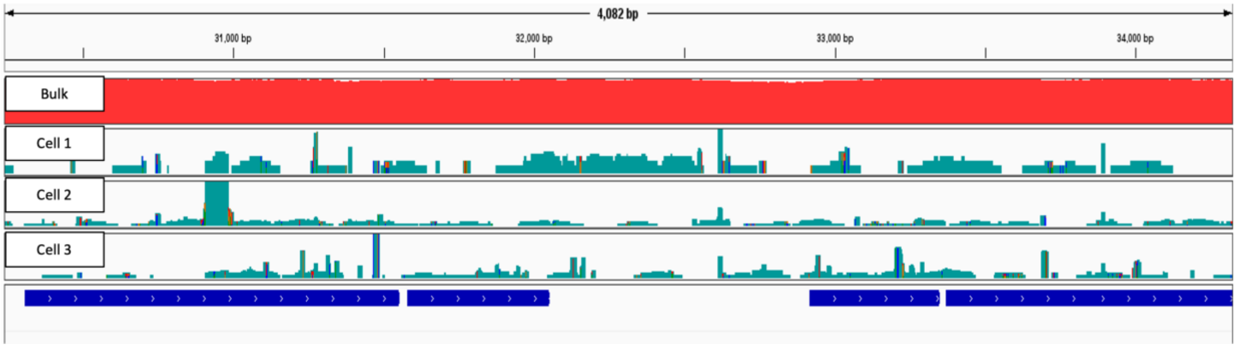
Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
hflX	**47.57912646**	**5.161361442**	**0.789037858**	**2.82E-12**	**1.98E-10**
dnaE	**235.808896**	**3.86661516**	**0.98851571**	**1.82E-11**	**9.28E-10**
fkpA	**95.64525324**	**7.397867815**	**0.89849135**	**1.21E-09**	**3.46E-08**
lon	**53.20681145**	**10.15983006**	**0.793041823**	**1.85E-08**	**4.42E-07**
rpoD	**12.49666404**	**12.50751631**	**0.701524755**	**2.02E-06**	**2.81E-05**
hslO	3.20376518	4.423146553	0.430277931	0.052037087	0.109414119
plsB	4.964760495	0.315877055	0.454382738	0.285925348	0.39815487
hslR	2.062975912	7.604030598	0.362224861	0.921967847	0.944083473

Appendix Table 7. DESeq2 result summary for type I CRISPR cas and enterobactin genes
Values were calculated based on differences between single-cell pseudo-bulk samples and true bulk samples. A negative log2FoldChange means that expression from single-cell pseudo-bulk is greater than bulk RNA samples. Bolded genes met the threshold (padj <0.05) for significant differential expression. baseMean is the mean normalized counts for all samples, lfcSE is the standard error, pvalue is the Wald test p-value, and padj is the Benjamini-Hochberg adjusted pvalue.

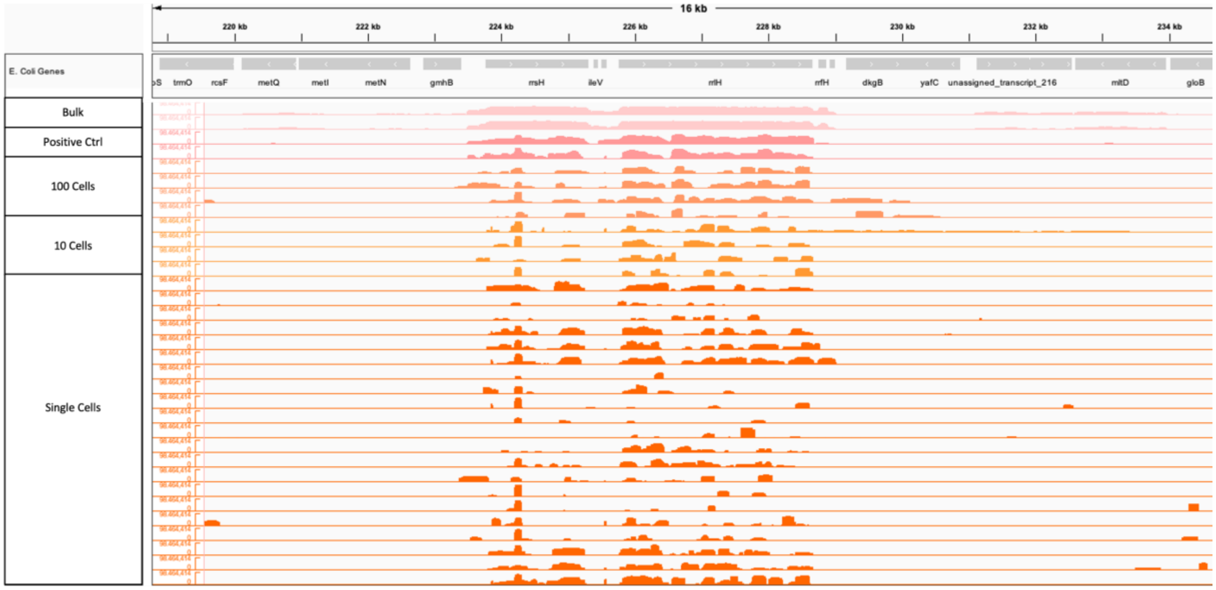
Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
entB	**3615.584612**	**-0.799121935**	**1.416219746**	**5.23E-08**	**6.79E-07**
entD	**58.33677862**	**-2.315417204**	**2.180586156**	**0.000452024**	**0.002262603**
fepA	**106.4934833**	**-1.988808775**	**0.825957821**	**0.001099924**	**0.004970915**
entF	80.19195744	0.890586928	0.540715237	0.05089504	0.085828913
fepE	19.15851415	-0.95850377	0.933932388	0.061650283	0.100797881
fepG	24.36688074	0.66371553	0.525769691	0.137120355	0.195961291
fepD	39.80590921	0.670360062	0.536789598	0.141492138	0.201123488
fepC	8.185493523	0.759801887	0.669851774	0.145813187	0.205917662
fepB	17.28781583	0.609865472	0.582704013	0.207938979	0.276685656
entA	41.0165339	0.465377172	0.477288068	0.260632573	0.333469286
entE	21.31795947	-0.438811904	0.624523658	0.350844416	0.425831923
fes	53.37347359	-0.446597515	0.632279283	0.358341517	0.433087017
entH	17.20393841	0.367339352	0.551401256	0.42077077	0.495535824
entS	31.81417345	-0.186402813	0.526958584	0.672428739	0.728677531
entC	8.726873075	0.133655108	0.62734564	0.785279572	0.825276527
casA	**437.1994473**	**-6.455892185**	**0.661741889**	**4.75E-23**	**1.27E-20**
casB	**914.2995786**	**-7.824619881**	**2.25683107**	**5.47E-10**	**1.21E-08**
cas3	**74689.35634**	**-0.947553862**	**1.588690265**	**2.03E-09**	**3.96E-08**
cas1	**414.1659298**	**-3.742865246**	**1.08740426**	**3.60E-06**	**3.03E-05**
casE	**55.43956242**	**-2.549800391**	**0.607607561**	**4.77E-06**	**3.82E-05**
casD	**49.80024123**	**-2.986479755**	**0.771877103**	**4.84E-06**	**3.86E-05**
casC	**34.45615799**	**-2.146073894**	**0.754585151**	**0.000399978**	**0.002033332**
cas2	13.26139494	-0.082759503	0.570806146	0.854374701	0.883835898



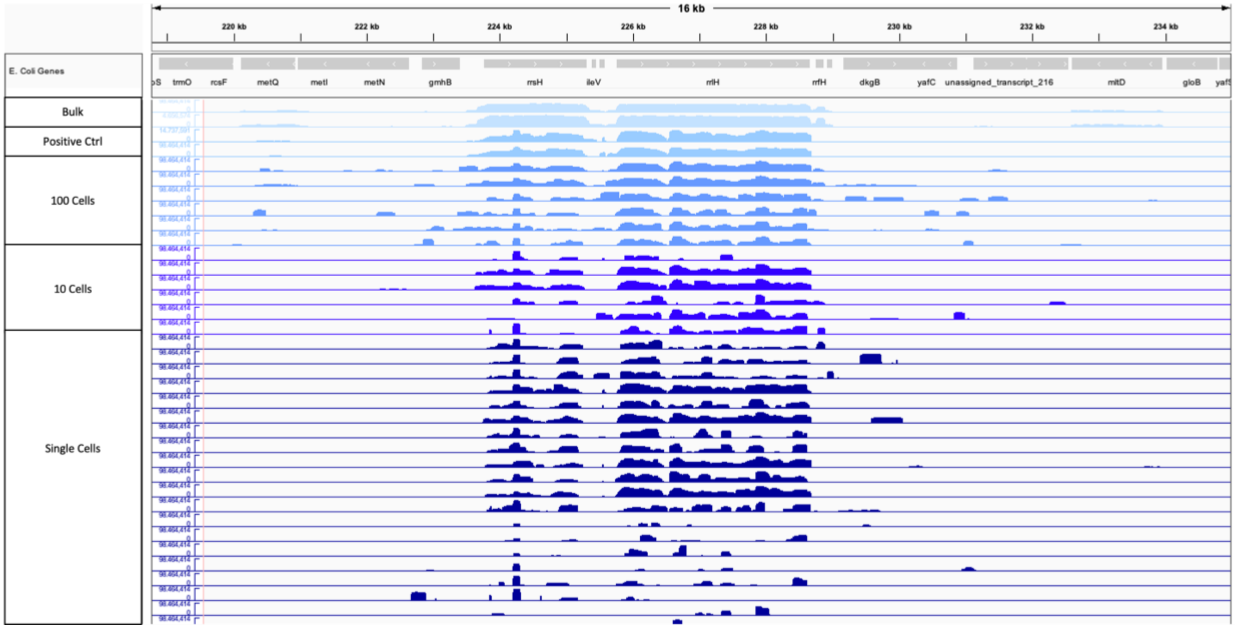
Appendix Figure 1. Genome coverage by sequencing effort 1.25 µL MDA reaction volumes
Trimmed reads were down-sampled from all 1.25 µL samples (n=5) to 200X, 100X, 80X, 60X, 40X, 20X using BBmap *reformat.sh*. Down-sampled reads were then mapped to *E. coli* MG1655 reference genome with *bbmap.sh*. Standard error bars were calculated using all five replicates.



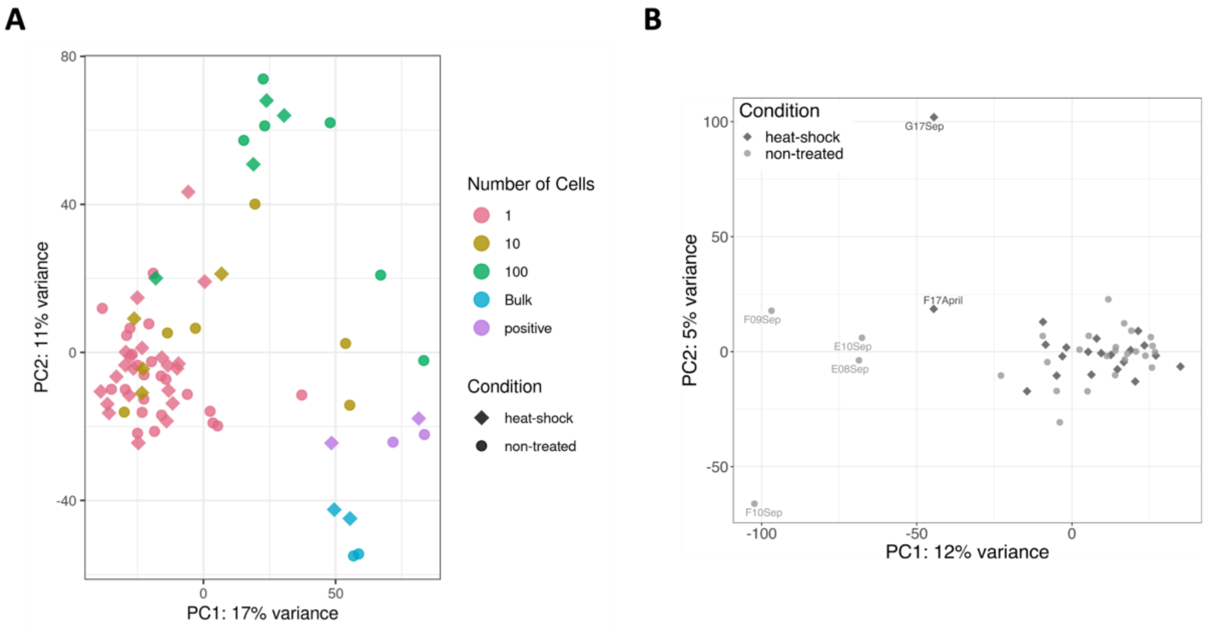
Appendix Figure 2. Read coverage and density plots for Liu et al (2019) scRNA-seq data
The very uniform and high coverage seen in the bulk RNA-seq data indicates DNA contamination. The consistent read mapping in intergenic regions for the single cells also indicates DNA contamination. Libraries were log scaled and view in Integrative Genomics Viewer.



Appendix Figure 3. Read coverage across the 16S rRNA operon for heat-shock samples
Non-uniform read mapping along the highly transcribed 16S rRNA region and in intergenic regions indicated that all libraries were unlikely to be DNA contaminated for heat-shock treated samples. Libraries were log scaled and view in Integrative Genomics Viewer.



Appendix Figure 4. Read coverage across the 16S rRNA operon for non-treated samples
Non-uniform read mapping along the highly transcribed 16S rRNA region and in intergenic regions indicated that all libraries were unlikely to be DNA contaminated for non-treated treated samples. Libraries were log scaled and view in Integrative Genomics Viewer.



Appendix Figure 5. PCA of the 300 most variant genes
A PCA of the 300 most variant genes between all samples. **B** Single-cell only PCA of the 300 genes most variant by condition.

7 Curriculum vitae

Morgan S. Sobol, M.Sc

Institute Address:
Hermann-von-Helmholtz Platz 1, B.601
76344, Eggenstein-Leopoldshafen, DE

morgan.sobol@kit.edu
Alt: morgan_starr_s@live.com
+49 0151 28889 2164

Education

Feb. 2019 – Present	**PhD in Microbiology and Biotechnology** Karlsruhe Institute of Technology *Advisor*: Prof. Dr. Anne-Kristin Kaster *Dissertation title*: Improving Single Cell 'Omics Methods for Investigating Microbial Dark Matter
Aug. 2016 - Aug. 2018	**Master of Science in Marine Biology** Texas A&M University-Corpus Christi *Advisor:* Dr. Brandi Kiel Reese *Thesis title*: Characterizing Novel Fungi from Oligotrophic Marine Deep Subsurface Sediments *Summa Cum Laude* Cumulative GPA: 4.0
Aug. 2013 - Aug. 2016	**Bachelor of Science in Biology**, minor in Chemistry Texas A&M University-Corpus Christi *Magna Cum Laude* Cumulative GPA: 3.884

Publications

1. **Sobol, M. S.** & Kaster, A.-K. Improving multiple displacement amplification for microbial single-cell genomics. (*Under review*).
2. Zoheir, A.E., **Sobol, M.S.**, Ordonez, D., Kaster, A.K., Niemeyer, C.M., & Rabe, K.S. A three-colour biosensor reveals multimodal stress response at the single cell level and the spatiotemporal dynamics of biofilms. (*Under review*).
3. **Sobol, M. S.**, Hoshino, T., Futagami, T., Kadooka, C., Inagaki, F. & Reese, B. K. Genome characterization of two novel deep-sea sediment fungi, *Penicillium pacificagyrus* sp. nov. and *Penicillium pacificasedimenti* sp. nov., from South Pacific Gyre subsurface sediments, highlights survivability. (*Under review*)
4. Kiel Reese, B.*, **Sobol, M. S**.*, Bowles, M. W. & Hinrichs, K.-U. Redefining the Subsurface Biosphere: Characterization of Fungi Isolated From Energy-Limited Marine Deep Subsurface Sediment. Front. Fungal Biol. 0, 49 (2021).
5. **Sobol, M. S.** & Kaster, A.K. Gezielte Zellsortierung in der Einzelzellgenomik. BIOspektrum 2021 273 27, 274–276 (2021)**
6. Kaster, A. K. & **Sobol, M. S.** Microbial single-cell omics: the crux of the matter. Appl. Microbiol. Biotechnol. 104, 8209–8220 (2020).

7. Dam, H. T., Vollmers, J., **Sobol, M. S.**, Cabezas, A. & Kaster, A.-K. Targeted Cell Sorting Combined With Single Cell Genomics Captures Low Abundant Microbial Dark Matter With Higher Sensitivity Than Metagenomics. Front. Microbiol. 11, 1377 (2020).
8. **Sobol, M. S.**, Hoshino, T., Futagami, T., Inagaki, F. & Reese, B. K. Draft genome sequences of Penicillium spp. From deeply buried oligotrophic marine sediments. Microbiol. Resour. Announc. 8, (2019).
9. Reese, B. K., Zinke, L. A., **Sobol, M. S.**, LaRowe, D. E., Orcutt, B. N., Zhang, X., Jaekel, U., Wang, F., Dittmar, T., Defforey, D., Tully, B., Paytan, A., Sylvan, J. B., Amend, J. P., Edwards, K. J. & Girguis, P. Nitrogen Cycling of Active Bacteria within Oligotrophic Sediment of the Mid-Atlantic Ridge Flank. Geomicrobiol. J. 35, (2018).

*denotes co-first authorship

** denotes non-peer-reviewed publication

Oral Presentations

July 2022	**Morgan Sobol,** Julia Münch, Benedikt Brors, Anne-Kristin Kaster. Improving single-cell RNA-seq for microorganisms. BIF-IGS Annual Retreat, Karlsruhe Institute of Technology. Eggenstein-Leopoldshafen, DE.
April 2021	**Morgan Sobol**, Anne Popova, Shraddha Chakraborty, Pavel Levkin, Anne-Kristin Kaster. Improving Multiple Displacement Amplification for Single-Cell Sequencing with Droplet Microarrays. CellME Tech Day Online: Single Cell Analysis.
March 2021	**Morgan Sobol**, Anne Popova, Shraddha Chakraborty, Pavel Levkin, Anne-Kristin Kaster. Improving Multiple Displacement Amplification for Single-Cell Sequencing with Droplet Microarrays. Virtual 4th Revolutionizing Next-Generation Sequencing Conference.
September 2019	**Morgan Sobol**, Hang Dam, John Vollmers, Angela Cabezas, Anne-Kristin Kaster. Targeted cell sorting combined with single cell genomics reveals novel Chloroflexi species. 4th Microbial Single Cell Genomics Workshop, Boothbay harbor, Maine, USA.
November 2018	**Morgan Sobol**, Tatsuhiko Hoshino, Fumio Inagaki, Brandi Kiel Reese. Analysis of two *Penicillium* genomes from the oligotrophic marine subsurface. Texas ASM Branch Meeting, Corpus Christi, Texas, USA.
June 2017	**Morgan Sobol,** Mayra Rodriquez, Tatsuhiko Hoshino, Fumio Inagaki, Brandi Kiel Reese. Characterization of Deep Marine Subsurface Fungi from South Pacific Gyre Sediments. AbGradCon 2017, Charlottesville, Virginia, USA.
March 2017	**Morgan Sobol,** Mayra Rodriquez, Tatsuhiko Hoshino, Fumio Inagaki, Brandi Kiel Reese. Growth Characteristics of Fungi from Oligotrophic Marine Deep Subsurface. Texas ASM Branch Meeting, New Braunfels, Texas, USA.
January 2017	**Morgan Sobol**, Mayra Rodriquez, Brandi Kiel Reese. Characterization of Deep Marine Subsurface Fungi from South Pacific Gyre Sediments. MARB Annual Retreat. Texas A&M University – Galveston, Galveston, Texas, USA.
December 2016	**Morgan Sobol**, Mayra Rodriquez, Brandi Kiel Reese. Characterization of Deep Marine Subsurface Fungi from South Pacific Gyre Sediments. 6th Annual MSGSO Student Research Symposium, Texas A&M University-Corpus Christi. Corpus Christi, Texas, USA.
December 2015	**Morgan Sobol**, Laura Zinke, Brandi Kiel Reese. Investigating the Differences in the Total and Active Community Structure of mid-Atlantic Ridge Sediment. 5th Annual MSGSO Student Research Symposium, Texas A&M University-Corpus Christi. Corpus Christi, Texas, USA.

Poster Presentations

February 2022	**Morgan Sobol**, Anne Popova, Shraddha Chakraborty, Pavel Levkin, Anne-Kristin Kaster. Improving Multiple Displacement Amplification for Single-Cell Sequencing with Droplet Microarrays. VAAMS 2022, Online, DE.
July 2021	**Morgan Sobol**, Anne Popova, Shraddha Chakraborty, Pavel Levkin, Anne-Kristin Kaster. Improving Multiple Displacement Amplification for Single-Cell Sequencing with Droplet Microarrays. BIF-IGS Annual Retreat, Karlsruhe Institute of Technology. Eggenstein-Leopoldshafen, DE.
February 2016	**Morgan Sobol**, Laura Zinke, Doug E. LaRowe, XinXu Zhanhe, Ulrike Jaekel, Fengping Wang, Beth N. Orcutt, Thorsten Dittmar, Heath J.Mills, Jan P. Amend, Katrina J. Edwards, Peter Girguis, Brandi Kiel Reese. Investigating the Differences in the Total and Active Community Structure of mid-Atlantic Ridge Sediment. ASLO 2016, New Orleans, Louisiana, USA.
October 2015	**Morgan Sobol**, Laura Zinke, Brandi Kiel Reese. Investigating the Differences in the Total and Active Community Structure of mid-Atlantic Ridge Sediment. 12th Annual Pathways Student Research Symposium, Texas A&M University-Corpus Christi. Corpus Christi, Texas, USA.
July 2014	**Morgan Sobol**, Emiko Sano, Scott Miller. Investigating functional divergence of *recA* paralogs from *Acaryochloris marina* in *Escherichia coli*. OREOS Undergraduate Research Symposium, University of Montana, Missoula, Montana, US

Awards and Scholarships

July 2022	Biointerfaces International Graduate School 1st Place Oral Presentation
December 2021	Karlsruhe House of Young Scientists Networking Grant
March 2020	Biointerfaces International Graduate School Travel Grant Award
May 2019	Deep Carbon Observatory Deep Life Community Travel Support Award
December 2017	TAMUCC Parent’s Council Travel Award
October 2017	Ruth A. Campbell Endowed Scholarship
September 2017	MARB Student Travel Award
May 2017	TAMUCC Parent’s Council Travel Award
2016-2017	TAMUCC College of Science and Engineering Graduate Scholarship
October 2015	Pathways Overall Second Place Undergraduate Poster Presentation
October 2015	Pathways 1st Place Undergraduate Life Sciences Poster Presentation
Summer 2015	LSAMP Undergraduate Research Award
2015 – 2016	Killebrew Scholarship Fund
2015 –2016	Rising Scholar Scholarship
2014 –2016	Marty Pritchett Scholarship Endowment
2013 –2016	Texas A&M University – Corpus Christi Academic Achievement Scholarship

Volunteer Work

July 2021	Biointerfaces International Graduate School Annual Retreat Organizer
October 2017	Texas State Aquarium Teen STEM Cafe
January 2017	2017 Coastal Bend Regional Science Fair Judge
2013- 2016	Texas State Aquarium, Corpus Christi, Texas